

THE
BRITISH JOURNAL
OF
STATISTICAL
PSYCHOLOGY

EDITED BY
CYRIL BURT

WITH THE ASSISTANCE OF
CHARLOTTE BANKS AND ARTHUR SUMMERFIELD

AND THE FOLLOWING EDITORIAL BOARD

A. C. AITKEN	L. S. PENROSE
M. S. BARTLETT	J. FRASER ROBERTS
W. G. EMMETT	A. RODGER
M. G. KENDALL	W. STEPHENSON
D. N. LAWLEY	P. E. VERNON
E. S. PEARSON	J. W. WHITFIELD

Managing Editor : H. G. MAULE

Volume VI
1953

UNIVERSITY OF LONDON PRESS LTD.

WARWICK SQUARE, LONDON, E.C.4.

Digitized by the Internet Archive
in 2026

A NOTE ON SIR CYRIL BURT'S 'FACTORIAL ANALYSIS OF QUALITATIVE DATA'

By LOUIS GUTTMAN

Scientific Director, Israel Institute of Applied Social Research, Jerusalem

I. *Procedures for Dealing with Qualitative Data.* II. *Analysis into Principal Components.* III. *Applications to Scalable and Non-Scalable Data.* IV. *The Study of Deviations : Image Analysis and Nodal Analysis.*

I. PROCEDURES FOR DEALING WITH QUALITATIVE DATA

World War II interrupted communication between scientists in different countries, and only gradually is the exchange of information being made up. A case in point is the monograph *The Prediction of Personal Adjustment* by Paul Horst and others, which appeared just before the United States actively entered the holocaust. This was published by the Social Science Research Council in New York as *Bulletin 48* in 1941. During the war, only a handful of copies was able to reach Europe. In a recent article in this *Journal* (1), Sir Cyril Burt has called attention to the importance of developing a theory for qualitative data, as distinct from quantitative data. He develops a particular algebraic formulation which leads to the resolution of the data into principal components (latent vectors). This happens to be a topic treated also in the above-mentioned monograph, by the present writer. It is gratifying to see how Professor Burt has independently arrived at much the same formulation. This convergence of thinking lends credence to the suitability of the approach. The purpose of the present note is to call attention to the points of similarity and to describe further developments which have occurred in the United States and in Israel in the theory of qualitative data.

In a chapter of the above-mentioned Social Science Research Council monograph entitled 'The Quantification of a Class of Attributes' (2), it was proposed that qualitative data could be recorded in a manner amenable to treatment by matrix algebra. In form, the matrix M on page 326 of the monograph is identical with Table I of Professor Burt's article ((1), p. 171). My own paper proposes three different kinds of problems of quantification :

(a) Deriving quantitative *scores* for the respondents, to maximize a certain correlation ratio (for the *rows* of the matrix).

(b) Deriving quantitative *weights* for the categories of the qualitative items, to maximize another correlation ratio (for the *columns* of the matrix).

(c) Deriving scores and weights simultaneously, to maximize a certain correlation coefficient (for rows and columns of the matrix simultaneously).

It was proved that all three problems are essentially equivalent. The maximizing scores and weights for the third problem are also the maximizing ones for the

first and second problems respectively. Furthermore, the maximum correlation ratios and coefficient are always mutually equal for the same data.

II. ANALYSIS INTO PRINCIPAL COMPONENTS

As usual in analysing internal consistency, the equations sought lead not only to a *maximum*, but also to a series of other stationary values down to a minimum. These various solutions provide a principal component system. There is a set of principal component scores for the respondents, and a set of principal component weights for the categories of the items. Professor Burt has arrived at the latter solution, namely, that for principal component weights.

My own article goes on to point out that, while the principal components here are formally similar to those for quantitative variables, nevertheless their interpretation may be quite different. The interrelations among qualitative items are not linear, nor even algebraic, in general. Similarly, the relation of a qualitative item to a quantitative variable is in general non-algebraic. Since the purpose of principal components—or any other method of factor analysis—is to help reproduce the original data, one must take into account this peculiar feature.

The first principal component can possibly fully reproduce all the qualitative items entirely by itself: the items may be perfect, albeit non-algebraic, functions of this component. *Linear* prediction will not be perfect in this case, but this is not the best prediction technique possible for such data. Therefore, if the first principal component only accounts for a small proportion of the total variance of the data in the ordinary sense, it must be remembered that this ordinary sense implies *linear* prediction. If the correct, but *non-linear*, prediction technique is used, the whole variation can sometimes be accounted for by but the single component. In such a case, the existence of more than one principal component arises merely from the fact that a linear system is being used to approximate a non-linear one. (Each item is always a perfect linear function of *all* the principal components taken simultaneously.)

III. APPLICATIONS TO SCALABLE AND NON-SCALABLE DATA

This thinking was carried further with the definition of a perfect scale (3). In a perfect scale, each item is a perfect function of a single rank order of respondents. The functional relation is not algebraic, although of an extremely simple nature. Applying the general equations of principal components developed in (2) to the case of the perfect scale defined in (3) leads to some most extraordinary results. There is a definite *law of formation* for the entire set of principal components for scores, and a corresponding law of formation for principal component weights. Indeed, the matrix equation for the principal components in each case transforms into a *difference* equation. For dichotomous items, this is a second order linear difference equation which, in the limit, for an infinite number of items, yields a second order *differential* equation that is classical in mathematics and physics. The principal components turned out to be sets of orthogonal functions, like the Legendre polynomials, Hermite polynomials, trigonometric functions, etc.

The second principal component in all cases is a perfect U- or J-shaped function of the first. The third principal component is also a perfect function of the first, but its graph shows a curve with two bends (for example, a cubic parabola). The

linear correlation between any two principal components is always zero, because they are orthogonal to each other (and have zero means) ; but the *curvilinear* correlation of each with the first component is perfect. In this sense, the further principal components beyond the first add no new information.

In another sense, it has been found that the further principal components provide a *psychological* frame of reference for scalable data. The second principal component for scores has been identified for many attitudinal data as the *intensity* of attitude (5). More recently, the third and the fourth component for scores have been found to portray somewhat new psychological concepts which have been named *closure* and *involution*, respectively (9).

For non-scalable data, the principal components will not, of course, have a law of formation such as that just described, and indeed may have no useful law at all. Progress has also been made in developing meaningful structural theories for non-scalable qualitative data, and here various new kinds of systems of principal components appear possible. For the beginning of these developments, which go under the general name of the 'theory of facets,' see (10).

An unexpected consequence of these discoveries for qualitative data has been the rethinking it has stimulated about factor analysis in general. A new approach has been developed for qualitative data with linear regressions, which also leads to laws of formation of principal components. The general theory is called that of the *radex* and it contains the *simplex* and *circumplex* as subtheories. Empirical data tend to confirm this new approach for the case of tests of mental abilities (11).

IV. THE STUDY OF DEVIATIONS: IMAGE ANALYSIS AND NODAL ANALYSIS

The above-mentioned type of difference equation, and its lawful principal components, holds for the case of a perfect scale. An important problem that must be considered arises from the fact that perfect scales are rarely, if ever, to be found in practice. More typical may be the case of quasi-scales, or scales 'plus' error. Here, then, it becomes important to distinguish between the structural and the deviant aspects of the observed data : only the former will have a law of formation.

Two new general approaches for dealing with the problem of deviations from perfect structures are called *image analysis* and *nodal analysis* respectively. Both of these are useful for many kinds of non-scale structures as well as for quasi-scales.

One image technique, Israel Alpha, for studying quasi-scales has been described in (14). The Israel Beta technique has also been used successfully, and will be described in future publications.

Nodal analysis can be regarded as a generalization and simplification of Lazarsfeld's *latent structure* theory. It leads to easy computing routines, whether for quasi-scale or non-scale structures (15).

Here we cannot do more than to call attention to these newer developments, most of which have not yet been published in much detail. The work done thus far in probing the possibilities of structural theories for qualitative data seems to indicate that we are merely scratching at the beginnings of a rich and intricate subject, which fully deserves to be studied in its own right.

REFERENCES

1. Burt, Cyril (1950). 'The factorial analysis of qualitative data.' *Brit. J. Psych., Stat. Sect.*, III, 166-85.
2. Guttman, Louis (1941). 'The quantification of a class of attributes,' in Horst, P., *et al. The Prediction of Personal Adjustment*. New York : Social Science Research Council, 319-48.
3. Guttman, Louis (1944). 'A basis for scaling qualitative data.' *American Sociological Review*, IX, 139-50.
4. Guttman, Louis (1947). 'The Cornell technique for scale and intensity analysis.' *Educ. and Psychol. Measurement*, VII, 247-80.
5. Guttman, Louis, and Suchman, E. A. (1947). 'Intensity and a zero point for attitude analysis.' *Amer. Sociol. Rev.*, XII, 57-67.
6. Guttman, Louis (1947). 'Scale and intensity analysis for attitude, opinion, and achievement,' in Kelly, George A. (editor). *New Methods in Applied Psychology*. Maryland : University of Maryland Press, 173-80.
7. Guttman, Louis (1950). 'The principal components of scale analysis,' ch. 9 in Stouffer, S. A., *et al. Measurement and Prediction*. Princeton : University Press.
8. Guttman, Louis (1951). 'Scale analysis, factor analysis, and Dr. Eysenck.' *Internat. J. Opinion and Attitude Research*, 5, 103-20.
9. Guttman, Louis. 'The principal components of scalable attitudes,' in Lazarsfeld, Paul F. (editor). *Mathematical Thinking in the Social Sciences*. New York : Columbia University Press (in the press).
10. Guttman, Louis. 'The theory of facets.' *Ibid.*
11. Guttman, Louis. 'A new approach to factor analysis : the radex.' *Ibid.*
12. Guttman, Louis, and Suchman, E. A. (1947). 'A solution to the problem of question "bias."' *Public Opinion Quarterly*, XI, 445-55.
13. Guttman, Louis, and Foa, Uriel G. (1951). 'Social contact and an intergroup attitude.' *Public Opinion Quarterly*, XV, 43-53.
14. Guttman, Louis (1951). *The Israel Alpha technique for scale analysis* (stencilled). Available on request from the Israel Institute of Applied Social Research.
15. Guttman, Louis. 'The theory of nodal analysis' (in preparation).

SCALE ANALYSIS AND FACTOR ANALYSIS

*Comments on Dr. Guttman's Paper*¹

By CYRIL BURT

University College, London

I. *The Analysis of Answer Patterns.* II. *Criticisms of the Factorial Approach.* III. *Factor Analysis of a Typical Scale Pattern.* IV. *Factor Analysis of the Corresponding Answer Pattern.* V. *Summary.*

I. THE ANALYSIS OF ANSWER PATTERNS

Quantitative and Qualitative Data. It is encouraging to find that Dr. Guttman has discerned points of resemblance between the methods we have independently reached for analysing qualitative data; and I willingly agree that this convergence in our lines of approach lends additional plausibility to the results. If, as I gather, he cannot wholly accept my own interpretations, that perhaps is attributable to the fact that our starting-points were rather different. My aim was to factorize such data; his to construct a scale. In the paper² to which he has referred, I sought to develop a practicable procedure for discovering the factors underlying qualitative characteristics, and then to apply that procedure to certain recurrent problems of my own. Dr. Guttman's purpose, as the title of his previous contribution indicates, was to present 'a theory and method of scale construction' by means of 'quantifying a class of attributes.' I have myself commented³ on the similarity between my equations for factor-measurements and the equations for component scores given by Dr. Guttman, and ascribed it to the fact that his mode of deriving components seemed identical with the method proposed by Pearson for calculating 'index characters'—a method which had formed the original basis for my own factorial procedures.

Dr. Guttman's technique of scale analysis has been developed primarily in connexion with "the study of the internal structure of a universe of attitude and opinion items"; and the data he has used to illustrate his procedure consist, as a rule, of responses to questionnaires. I myself, however, first encountered the type of problem which he has discussed in a rather different field, namely, in the attempt to refine the scaling of the original Binet-Simon intelligence-tests. Before the publication of Binet's 1911 version, British psychologists relied chiefly on tests of an 'internally graded' type—opposites, dotting, card sorting, and the like. Binet introduced a composite *échelle* comprising a number of 'externally graded items,' i.e., questions or tasks for which the responses differ not in quantity but only in quality, and which are simply marked 'right' or 'wrong.'⁴ Thus,

¹ I am much indebted to Dr. Guttman for his kindness in reading my manuscript and to Dr. M. L. Fraser for his help in checking my calculations.

² This *Journal*, III, pp. 166–85. As there explained, the theoretical part of the paper was based on a technical memorandum prepared during the early stages of the war for the Department of Personnel Selection of the War Office. The factorial procedure had already been applied to qualitative physical characteristics, and to data obtained from the Binet-Simon Scale, for which it was originally developed.

³ This *Journal*, V, p. 200. Dr. Guttman observes (p. 2 above) that the method I have followed yields the same solution as his for the 'principal component weights.' But, as I shall show in a moment, there is (with a suitable standardization of the data) a still more striking resemblance between the figures reached by the two different methods for the 'principal component scores.'

⁴ For a discussion of the principles underlying the two types of test, see Burt, C. (1914), 'The Measurement of Intelligence by the Binet Tests,' *Eugen. Rev.*, VI, pp. 36–50, 140–52.

with an internally graded test the numerical data consist of measurements ; with the externally graded, of frequencies. Hence for the former the appropriate statistical methods appeared to be those developed for 'quantitative variables' (e.g., the ordinary product-moment correlation) ; for the latter, those described by Yule for what he called 'attributes' (e.g., methods of contingency and association).

In the chapters contributed to *Measurement and Prediction* (18) both Dr. Guttman and Dr. Lazarsfeld draw a sharp distinction between the principles involved in these two cases. Factor analysis, they maintain, has been elaborated solely with reference to data which is quantitative *ab initio* ; hence, they suppose, it cannot be suitably applied to qualitative data. On this side of the Atlantic, however, there has always been a tendency to treat the two cases together, and, with this double application in view, to define the relevant functions in such a way that they will (so far as possible) cover both simultaneously.¹ British factorists, without specifying very precisely the assumptions involved, have used much the same procedures for either type of material. Nevertheless, there must of necessity be certain minor differences in the detailed treatment. These were briefly indicated in the paper Dr. Guttman has cited ; but they evidently call for a closer examination. I think in the end it will be found that they are much slighter than might be supposed.

Example. With qualitative material, the special problems that arise and the general mode of approach can be best explained if we start with a concrete illustration. Suppose we want to classify the children in some particular age-group—all those aged six, let us say—into 5 grades of ability by means of a series of dichotomous test-items like the questions in the Binet Scale. If each item constituted a pure and reliable test for the ability to be assessed, then, arguing along the lines of 'information theory,'² and proceeding by a progressive subclassification, we should evidently need six items altogether : namely, (i) a test like 5 Numbers, which would separate those below average from those above, i.e., those with a mental age of less than 6 from those with a mental age of 6 or more ; (ii) a test like 4 Numbers, which would divide those with a mental age below average into those with a mental age of less than 5, and those with a mental age of 5 to 6 ; and (iii) a test like 3 Numbers, which would divide those with a mental age of less than 5 into those with a mental age of less than 4, and those with a mental age of 4 to 5. Similarly, we should require three more tests to subclassify those who were above the average for age 6.

In an experiment planned to study the working of the tests, the data obtained would be entered on a mark sheet similar to that shown in Table I. This is a special case of the more general table³ printed in my previous article ((17), Table I). Now, however, each 'determinable' presents two 'determinates' only, i.e., each item yields two 'categories,' namely, 'pass' (marked 1) and 'fail' (marked 0). British investigators⁴ describe such tables as 'answer patterns' ; and, from the early days of mental testing, they have devoted a good deal of attention to the simpler types of pattern which the answers from these non-graded tests appear to display.

Table I records the results obtained with six tests from the London version of the Binet Scale, when applied to a representative sample of 6-year-old school children, from whom, for simplicity, I have chosen 10 individuals only. It may be compared with the typical table printed by Dr. Guttman in his discussion of the 'components of scale analysis' ((18), p. 317, Table I).

The method of recording differs from his in two minor respects. First of all, in accordance with the principle followed by most practical test-constructors, the battery is designed to contain at least one item so easy that all the children can answer it and at least one so difficult that all are

¹ See, for example, (14), I, pp. 15f., where the Stieltjes integral is used to facilitate this combined treatment. The work of the French school (Borel and his colleagues), by deriving the theory of measurement from the theory of sets, makes the relations still simpler : (see (15), opening chapters). A yet more recent and relevant example is Kendall's 'general coefficient of correlation' which applies equally well to measurements, ranks, and dichotomous classifications (cf. Kendall, M. G., *Rank Correlation Methods*, 1948, pp. 17f.).

² Cf. Burt, C., this *Journal*, IV, p. 196. The principles involved are analogous to, but of course not identical with, the conversion of an integral measurement from the 'common scale of notation' to the 'binary' where only two digits, 0 and 1, are used.

³ Dr. Guttman says in his paper (p. 1 above) that this table is identical in form with "the matrix *M* on p. 326 of the monograph." I should have thought it was rather to be compared with the very first table in his monograph (p. 321), since that (like mine) presents the initial record sheet which we require to analyse, and shows the various items explicitly classified into categories and subcategories.

⁴ The same mode of tabulation is adopted by Ferguson and other British writers : cf. (6) and (11).

TABLE I. ANSWER PATTERN FOR DICHOTOMOUS TESTS

Tests	Persons										Total	Differences
	A.A.	B.A.	B.B.	C.A.	C.B.	C.C.	C.D.	D.A.	D.B.	E.A.		
Repeating—												
3 Numbers..	1	1	1	1	1	1	1	1	1	1	10	1 2 4 2 1
4 Numbers..	0	1	1	1	1	1	1	1	1	1	9	
5 Numbers..	0	0	0	1	1	1	1	1	1	1	7	
6 Numbers..	0	0	0	0	0	0	0	1	1	1	3	
7 Numbers..	0	0	0	0	0	0	0	0	0	1	1	
8 Numbers..	0	0	0	0	0	0	0	0	0	0	0	
Total ..	1	2	2	3	3	3	3	4	4	5	30	
Deviates ..	-2	-1	-1	0	0	0	0	+1	+1	+2		
Frequencies ..	1	2		4				2		1	10	

likely to fail. Hence $N + 1$ tests are deemed necessary to discriminate N types of persons. For theoretical work, however, where we are more interested in factorizing items than in grading persons, it is better to include $n + 1$ types of persons in order to analyse the answers from n tests. Secondly, the ordinary answer pattern gives only the positive replies, whereas Dr. Guttman enters negative as well as positive, and thus has $2n$ categories for n tests. It will be convenient to keep his phrase—'scale pattern'—to designate his type of table, and to reserve the phrase 'answer pattern' to designate a mark sheet of the more familiar type.

Problem. In the earlier investigations of the Binet and similar tests, one of my chief objects was to determine the factors in terms of which such answer patterns could best be described. The factors appeared to be of two kinds—'content factors' and 'difficulty factors'.¹ Could we so purify our scale that it was virtually a perfect instrument for measuring 'general intelligence,' all factors resulting from special abilities would be eliminated. In such a case there would still remain two obvious factors for which we should wish to secure specifications—the general factor of 'difficulty,' common to all the tests, and the general factor of 'ability,' common to all the persons.

In practice the former is usually obtained by correlating persons, the latter by correlating tests. If we insist on calculating correlations, we must begin by reducing the crude scores to deviations. But if this preliminary reduction is omitted, both factors can be obtained from the same matrix by factorizing rows for the one and columns for the other. With dichotomous items, however, like those illustrated in Table I, provided we are content to give each person and each item an equal weight, we can avoid the more elaborate calculations required for a full and formal analysis, and assess the difficulty of the items by simply adding the number of correct replies obtained for each item, and the ability of the children by simply adding the number of correct replies given by each child: (cf. the marginal totals in Table I). As the answer pattern above plainly indicates, there is a certain reciprocity between the two factors; and, indeed, either concept may be defined in terms of the other.²

The Unique Answer Pattern. When the sample of items and the sample of persons have been picked more or less at random, then the frequency-distributions for both will be approximately normal or at least bell-shaped. In practice, however, an attempt is generally made to choose items that will yield a rectilinear scale. Binet's method of standardizing his problems by grouping children according to chronological age was intended to provide a rectilinear distribution for persons as well (or rather for categories of persons). Evidently, if we had one perfectly reliable item for each level of difficulty and one perfectly representative child for each level of ability, the resulting answer pattern would then be uniquely determined.

Let us consider the effect of such a double selection of items and persons on an answer pattern like that set out above. When the items have been efficiently selected, but the children are unselected (or just randomly selected from each age-group), then the line dividing successes from failures will approximate to a normal ogive. In such a case, if the number of children is large, the table showing the answer pattern is usually condensed so as to allot a column to each age-group

¹ The 'difficulty' factors found in such analyses have been criticized by Mr. Gourlay as artefacts. For his criticisms and a reply, see this *Journal*, IV, pp. 65–76.

² Cf. Burt, C., 'Test Construction and the Scaling of Items,' this *Journal*, IV, 1951, p. 106, Sect. III (a).

instead of to each child, and percentages are usually entered in place of separate units ; but these percentages still exhibit an approximation to the ogival form (cf. (4), pp. 132-3, Table III, and (19), pp. 113f.). When both the tests and the persons have been appropriately selected, the dividing line should be perfectly straight, and the area containing the successful answers a right-angled triangle.

From this it will be clear that the peculiar type of answer pattern which would be given by a test comprised of a perfectly consistent set of items would possess several distinctive peculiarities.¹ (i) If with such a pattern of answers we compute an order of difficulty for the test-items from any sub-group of persons within the sample, that order should always remain the same, no matter how the sub-group is selected. (ii) The N_i persons passing any given item will always be the N_i most able persons in the group, no matter what N_i may be. (iii) A total of x marks will always be obtained by passing the x least difficult items in the series, no matter what x may be. It follows (iv) that the frequencies in the distribution of persons (shown near the bottom of the table) will be identical with the first differences between the successive numbers of persons passing each test (shown at the side of the table). Thus, as Walker points out (6), except so far as there may be ties for individuals having the same level of ability, or for items belonging to the same level of difficulty, the answer pattern will be completely fixed by the order of the items and the order of the persons. If, therefore, the frequency distribution of the persons is stated, we can deduce from the total mark gained by any one person, not only how many items, but also which items, he has in fact passed.

The Effects of Inconsistency. With the majority of the pass-or-fail tests in common use, the order of difficulty is far from stable ; and the resulting pattern never displays the perfect regularity just described. The Binet tests, for example, are intended to measure a hypothetical factor called general intelligence : how far is it safe to assume that all the items consistently measure this factor and nothing else ? With the six memory tests enumerated above, any child who answered a harder test would almost certainly answer all the easier tests, and any child who failed in an easier test would almost certainly fail in all the harder. But now suppose we add Binet's question requiring the child to give the date. A bright child of 6 who scores a mental age of 8 will probably fail in the new test but succeed in repeating 6 numbers ; a dull child of 10 with the same mental age will probably fail with 6 numbers but succeed in the new test. Omit '6 Numbers,' and the second will seem the more intelligent ; omit 'Giving the Date,' and the first will score most. Hence the verdicts of the two tests are no longer consistent ; and, largely for this reason, Binet dropped both from the final version of his scale.²

For such irregularities there are two main causes. Both persons and items suffer from what might be called an internal unreliability. To some extent this can be obliterated by grouping. We may, for example, begin by grouping persons according to age, ability, sex, social status, and the like. Even then we usually find that the order varies for different groups : (cf. the tables of rank-differences in (4), pp. 143, 194). When they are large enough to be statistically significant, such variations imply (in the language of analysis of variance) an 'interaction' between the special abilities or special attainments of the persons and the special content factors of the tests. For certain statistical purposes, the differences can be still further smoothed out by grouping the items. But for practical purposes it is better to recognize their existence, and regard them as a cue for improving the way the items have been selected. Binet, indeed, in his successive revisions attempted to purify his scale by eliminating most of those that appeared to be affected by scholastic attainments or by special aptitudes. However, he neither used nor possessed any objective method for determining which items were relatively pure and which were not.

A theoretical study of the structures exhibited by perfect and imperfect scales shows that irregularities of the type described must inevitably lower the efficiency of the test-series as a whole, and explains just how such irregularities operate. In particular it is found that they tend (a) to diminish the validity of the separate items, as assessed by the correlation of each with the total marks obtained from all, and (b) to render the composite test less reliable, less valid, and (since they inevitably reduce the variance of the marks) less discriminative (cf. (11), pp. 55-8). It therefore becomes desirable to find some method for measuring what I have called the 'homogeneity' and 'heterogeneity' of any test-battery proposed for practical use.

¹ Ferguson's account of the properties of the 'unique answer pattern matrix' ((11), p. 53) seems incomplete. As an earlier critic has pointed out, "he begins with 'firstly,' but then either he (or the printer) appears to have omitted certain characteristic properties, specified by other writers." Nor can I agree with Ferguson's suggestion that the "unique configuration" furnished by a perfect scale "seems to indicate that the mind has a certain structure" (p. 59) : it results from the way the investigator has constructed his composite test.

² A number of similar inconsistencies between the responses obtained with the original version of the scale were pointed out in an early report. The most conspicuous occurred with semi-scholastic items like Transcription and Reading—tasks which nearly every normal child learns to perform at a fairly definite age in his school life, whether he is slightly above or slightly below the average level of intelligence. The irregularities in the actual answer patterns obtained from the earlier surveys furnished the first and most obvious indications of the presence of such unsatisfactory items : (cf. (4), pp. 132-4, 146f.).

The Measurement of Heterogeneity. The general characteristics of consistent and inconsistent answer patterns will now be clear. As we have seen, when the tests are arranged in order of difficulty and the persons in order of ability, then the answer pattern for an ideal test should display a solid block of 1's above the diagonal line of division (i.e., in what I shall call the 'success triangle') and a solid block of 0's below it (in what I shall call the 'failure triangle'). Inconsistencies will be indicated by erratic 0's appearing in the former area and erratic 1's in the latter; and the pattern will in consequence no longer show a sharp or cleanly defined line separating all the individual passes from all the individual failures. The effect of these irregularities Thomson designated by the term 'higgledypiggledyness.' Any coefficient obtained by counting up their number (or some equivalent procedure) he proposed to call a 'coefficient of hig' and its complement a 'coefficient of unig' (uniqueness). The words he has coined avoid the ambiguities of vaguer terms, like 'heterogeneity' and 'inconsistency,' which are used in so many contexts with so many different meanings.

With a perfectly haphazard distribution of answers we should expect half the total number of successes to be scattered over one triangle, and half over the other. An obvious index of homogeneity or 'unig' (U) is therefore given by the formula

$$U = \frac{\text{No. of Actual Successes in Success Triangle} - \frac{1}{2} \text{ Total No. of Successes}}{\frac{1}{2} \text{ Total No. of Successes}}$$

Writing this in the form

$$U = 1 - \frac{4 \times \text{No. of Successes in Failure Triangle}}{\text{Total No. of Responses}} = 1 - \frac{4 \sum e_f}{n \times N},$$

we see that it is virtually the same as Guttman's 'coefficient of reproducibility' ((18), p. 77).

In practice, however, with the ordinary type of composite test, such a coefficient does not prove altogether satisfactory. For one thing, with the same set of items, it varies according to the distribution of the sample of persons. In such a case an alternative method is to compare the frequencies for the individuals with the first differences between the totals for the tests; and accordingly, after considering various possibilities, Walker proposed a 'coefficient of hig' based on the sum of the squares of the discrepancies between the two sets of figures. But even this, as he himself points out in a more recent paper, is not entirely free from objection (6).

There are several other ways in which a coefficient might be conceivably developed. We might, for instance, use any of the coefficients of intrinsic reliability, e.g., the split half method, or the variance method which I have described as giving the mean value for all possible split halves.¹ Or again we might construct from the data the ideal answer pattern which gives the closest fit, and test the discrepancies by the square contingency ('chi-squared'); the agreement could then be measured by the root mean square contingency ϕ . When the scales or questionnaires include a large number of items, such procedures become decidedly laborious; and probably most British writers would concur with Ferguson's conclusion: "at present no convenient measure is available for estimating the divergence of an obtained answer-pattern matrix from a theoretically unique matrix" ((11) p. 54).

The Advantages of a Factorial Analysis. For general purposes, the best course, in my own view, is to undertake a factorial analysis of the data obtained. Since we are interested primarily in difficulty factors, the natural approach would be to correlate persons. If the persons are grouped in categories, the matrix of correlations or product-sums need not be large. Indeed, it will often be sufficient to compare the rank-orders of the several items for the different groups ((4), pp. 132f., 146f.). For more intensive studies, a complete analysis is desirable,² particularly as that may incidentally shed light on the presence of

¹ Burt, C., 'The Reliability of Teachers' Assessments,' *Brit. J. Educ. Psychol.*, XV, 1945, pp. 91-2. The formula giving the mean for all possible split-halves appears to be essentially the same as that given by Kendall for the 'coefficient of concordance' ((14), p. 411).

² This is especially true where the data relate to non-cognitive characteristics, e.g., social attitudes, Rorschach answers, and the like. Many of the items in a case-history are as a rule recorded only in terms of presence or absence—e.g., the occurrence of some accident, disease, or hypothetical gene, and especially certain recurrent features in the social environment—parents or relatives of this or that type, membership of this or that group or circle, contact with this or that member of the group. In all such cases the resulting 'nought-and-one' record-table can be treated like a 'measurement matrix' and factorized in the manner described; and the results obtained often yield suggestive ways of describing the intricate pattern of a life history or a social group (cf. (17), p. 166 and refs.).

irrelevant content factors. In such a case it will be better to start by correlating or comultiplying items rather than persons. The method of weighted summation, with unreduced item-variances, should be employed by preference. But simple summation will furnish a good approximation. The working procedures, and the results obtained in the case of empirical test-items, have been sufficiently discussed and illustrated in earlier papers (cf. (13), (15), (17), and refs.). In view of recent criticisms, however, it seems desirable to consider more closely how far it may be legitimate to treat qualitative data in this fashion, and particularly whether such methods can yield satisfactory results in the case of a perfect scale.

II. CRITICISMS OF THE FACTORIAL APPROACH

The Relations between Scale Analysis and Factor Analysis. In his very instructive paper (pp. 1-4) Dr. Guttman has indicated what he takes to be the differences, as well as the similarities, between his approach and my own. He observes that, while the 'principal components' employed in his method of scale analysis may be 'formally similar' to those employed in factor analysis, "nevertheless their interpretation may be quite different." The differences are examined more fully in his earlier contribution on 'The Relation of Scalogram Analysis to Other Techniques.'¹

If I understand Dr. Guttman rightly, his objections turn on five main points.

1. First, he holds that factor analysis is "designed only for quantitative variables," and is consequently unsuited for qualitative data ((18), p. 191). The method, he says, "originated as a single factor theory by Spearman, and was developed into a multifactor theory by Thurstone and others." Were his own procedure to be described in factorial terms, then, he adds, we should have to treat it as "a single factor theory for qualitative data."

However, as other writers have pointed out (cf. this *Journal*, V, p. 206), such statements limit the term 'factor analysis' to very specific forms. In point of fact, the technique now generally known as factor analysis is much older than Spearman's procedures. It originated with Pearson's proposal to reduce a given multivariate distribution to terms of the 'principal axes of the frequency ellipsoid' and take these axes as representing 'index characters.' It was Pearson's investigation of the general problem that really supplied the earliest "algebraic formulation, leading" (if I may borrow Dr. Guttman's phrase) "to the resolution of the data into principal components (latent vectors)" ((2), pp. 559f.; cf. (15), p. 309). Factor analysis was thus a multifactor method from the start. Spearman's single factor method was developed several years later as a substitute, because he held that Pearson's approach was unsuited to the data obtained in psychology. Nevertheless, in spite of his criticisms, numerous researches were carried out in which Pearson's method was applied, not only to quantitative measurements, such as those furnished by graded tests, but also to qualitative data, such as were supplied by dichotomous test-problems and by questionnaires. In a footnote, Dr. Guttman ((18), p. 193) refers to the treatment worked out by Yule (Karl Pearson's assistant) for dealing with frequency-tables for qualitative variables (3), and considers it "strange that statistical text-books in the social sciences have not followed suit, but fail to discuss material of this kind at all." But as a matter of fact in this country psychologists have made free use of Yule's procedures, especially in relation to social data²; and numerous theses could be cited where factorial methods have been applied to tables of frequencies, contingencies, or Yule's coefficients of colligation or point-correlation. In particular, 'answer patterns' have regularly been subjected to a factorial analysis by various devices (cf. (6), p. 326, (11), p. 52, and refs.).

¹ ((18), pp. 191f.: in the same volume Dr. Lazarsfeld also brings forward somewhat similar criticisms; cf. *ibid.*, pp. 469f.).

² In dealing with the item-responses obtained with the Binet Scale (where the classification was in terms of pass or fail) and with many social conditions affecting delinquency (where the classification was in terms of presence or absence) Yule's formulæ were used at a very early stage (cf. *Mental and Scholastic Tests*, 1921, pp. 217f., and *The Young Delinquent*, 1925, pp. 54f.). The tetrachoric coefficient advocated by Pearson and criticized by so many recent writers came into vogue somewhat later (cf. (13)). Yule's method of treating frequency-tables in cases of manifold classification involves fitting the matrix of observed frequencies with a matrix of expected frequencies derived by a procedure identical with that of centroid analysis. And when, instead of the more laborious method of 'weighted summation' entailed by Pearson's equations, I put forward the formula for 'simple summation'—the formula afterwards termed the 'centroid formula' by Thurstone—I pointed out that it was, in effect, an extension to quantitative variables of the usual formula for investigating association in the case of qualitative attributes (cf. (3), p. 64, eqns. (1) and (2)).

2. Secondly, Dr. Guttman argues that, in order to apply factor analysis, we must begin by calculating correlation coefficients, and that in the case of qualitative data such coefficients are bound to be misleading. With his criticisms of the uses made of the tetrachoric coefficient and the point correlation I very largely agree. Yet his arguments seem only to prove that these coefficients are not suitable for *all* occasions or for *every* purpose. There is one coefficient which he does not explicitly discuss—the ordinary product-moment coefficient applied to the data after they have been transformed to standard measure; and this, which has been frequently used for such problems, yields (as I shall show in a moment) results that are virtually the same as his. Nevertheless, nothing in the theory of factor analysis confines its application solely to coefficients of correlation. Indeed, for the factorist there is often a special advantage in working with frequencies, since (as I pointed out in the paper to which Dr. Guttman refers) the higher order frequencies may prove especially serviceable in the calculation of group factors.¹

3. His third argument runs as follows. The principal criterion for scalability is reproducibility. But factor analysis does not allow us to reproduce the original data from the so-called factor-measurements. Hence factor analysis can never show whether a scale is perfect or not. This objection, however, applies only to the special procedures advocated by Spearman and Thurstone, which seek to analyse, not the total variance, but merely the common factor-variance. The method of principal axes, on the other hand, requires the full test-variances to be retained in the covariance matrix; and with Pearson's procedure the factor-measurements are obtained by pre-multiplying the initial measurements by the matrix of direction cosines (the latent vectors). Now such a matrix is necessarily orthogonal. Hence its transpose can be used as a second pre-multiplier to reproduce the initial measurements from the factor-measurements (cf. (10), Appendix II). An exact reproduction is therefore possible. If, then, we can also show that, with a perfect scale, one of the factors so obtained is in perfect correlation with the rank of the persons, it would seem that the method can after all provide an entirely satisfactory criterion.

4. "The Spearman-Thurstone approach to factor analysis," as Dr. Guttman says, "is completely linear, and is therefore not adequate for analysing the curvilinearities inherent in the scale pattern." Certainly in that mode of approach the factor-measurements are always estimated by the method of linear regression, as developed in Pearson's earlier papers. But Pearson himself also elaborated a method for dealing with curvilinear regressions.² His treatment was intended primarily for problems involving an external criterion; but it is equally applicable to the case of an internal criterion or factor. Elsewhere I have argued that it is quite unnecessary to restrict the theory of factor analysis to linear relations only ((10), p. 258); and, with the aid of the orthogonal polynomials used in the theory of curvilinear regression,³ it is simple to estimate factor-measurements from test data or test data from factor-measurements on the assumption of non-linear relations.

5. Finally, Dr. Guttman concludes that "from a scale analysis it can be known what a factor analysis will show; from a factor analysis it will usually be difficult, if not impossible, to know what a scale analysis will show." To determine this point I propose to apply a factorial procedure to his own table, and see how far the results achieved are similar to those reached by his own scale analysis. A concrete numerical example will probably help best to explain how far our methods are similar and in what ways they seem to differ.

III. A FACTOR ANALYSIS OF A TYPICAL SCALE PATTERN

An Answer Pattern for Five Items. In Table II I have reproduced the 'scale pattern' which Dr. Guttman uses in his chapter on 'The Principal Components of Scale Analysis' ((18), pp. 317, 333) to illustrate his own procedure. It represents a 'perfect scale' formed by the responses of 6 'types of person' to five test-items classified under 10 'types of category': (for brevity I shall refer to them simply as 'persons' and 'categories'). In the last line I have appended the value to which the entries should be changed in order to secure the same square-sum (1·000) for each category.

¹ (17), p. 181. The memorandum already quoted gave a worked example from the Binet tests, and an algebraic proof (summarized in this *Journal*, V, p. 202).

² The earliest systematic study of curvilinear regression is to be found in Pearson, K., 'On the General Theory of Skew Correlation and Non-Linear Regression,' *Biometric Series*, II, 1905; cf. also *id.*, 'On a General Method of Determining the Successive Terms in a Skew Regression Line,' *Biometrika*, XIII, 1921, pp. 296f.

³ See (14), II (esp. pp. 158f., 'Curvilinear Regression: Case when the Independent Variate proceeds by Equal Steps,' and refs.), (5), and (8). Greenleaf (7) also includes useful refs.

This table corresponds to Table I in my previous article ((17), p. 171). There, however, the possibility was envisaged of more than two categories for each item, and those referring to the same item were kept together. Here there are two categories only—positive (e.g., correct answers) marked *a*, and negative (e.g., incorrect answers) marked *b*. Adopting Dr. Guttman's mode of tabulation, I have now grouped all categories of the same type together, regardless of item, and have numbered the items from 1 to 5 as their difficulty decreases, and the persons from 5 to 0 as their ability decreases. The reader may perhaps find it easier to interpret the factorial procedure if he thinks of the items as five test-problems from an ideal version of the Binet Scale, applied to typical children of six successive mental ages.

TABLE II. SCALE PATTERN FOR FIVE DICHOTOMOUS ITEMS (GUTTMAN)

Persons	Categories										Frequency of Response
	1a	2a	3a	4a	5a	1b	2b	3b	4b	5b	
5	1	1	1	1	1	0	0	0	0	0	5
4	0	1	1	1	1	1	0	0	0	0	5
3	0	0	1	1	1	1	1	0	0	0	5
2	0	0	0	1	1	1	1	1	0	0	5
1	0	0	0	0	1	1	1	1	1	0	5
0	0	0	0	0	0	1	1	1	1	1	5
Frequency of Response	1	2	3	4	5	5	4	3	2	1	30
Standardized Value ($1/\sqrt{N_j}$)	1	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{3}}$	$\frac{1}{\sqrt{4}}$	$\frac{1}{\sqrt{5}}$	$\frac{1}{\sqrt{5}}$	$\frac{1}{\sqrt{4}}$	$\frac{1}{\sqrt{3}}$	$\frac{1}{\sqrt{2}}$	1	

Factor Analysis. The method of weighted summation involves analysing the symmetric matrix of standardized product-sums into its latent roots and latent vectors (cf. Burt, *Factors of the Mind*, Appendix II). Accordingly we begin by calculating a symmetric contingency table, i.e., the matrix of product-sums for items, $C_i = M_i M_i'$ ((17), p. 172, eqn. (i)); and then standardize¹ it in the usual way by pre- and post-multiplying it by a diagonal matrix, $D^{-\frac{1}{2}}$, where the elements of $D^{\frac{1}{2}}$ are $\sqrt{N_j}$ (where N_j denotes the number of persons in each category). The figures thus obtained are shown in Table III: (to save space I omit the last 4 rows, since they are obvious from the fact that the complete table must be axisymmetric).

TABLE III. STANDARDIZED MATRIX OF PRODUCT-SUMS (*M*)

Category	1a	2a	3a	4a	5a	1b	2b	3b	4b	5b
1a	1.000	.707	.577	.500	.447	.000	.000	.000	.000	.000
2a	.707	1.000	.816	.707	.632	.316	.000	.000	.000	.000
3a	.577	.816	1.000	.866	.775	.516	.289	.000	.000	.000
4a	.500	.707	.866	1.000	.894	.671	.500	.289	.000	.000
5a	.447	.632	.775	.894	1.000	.800	.671	.516	.316	.000
1b	.000	.316	.516	.671	.800	1.000	.894	.775	.632	.477
..	..	..	..	..	..	..	..	..	..	..

(a) *Factor-Saturations.* The 'saturations' for the *i*th factor have been computed in the usual way by multiplying the normalized elements specifying the *i*th latent vector by the square root of the *i*th latent root. Six factors are obtained; and the resulting matrix is orthogonal by columns (Table IV).

¹As explained in my previous article, the main object of standardizing the data is to weight the frequencies so as to compensate for the differences in numbers in the various sub-categories. Here an incidental effect is to weight each item according to its difficulty.

TABLE IV. FACTOR-SATURATIONS FOR CATEGORIES (F)

Types of Category	Factor	G	BI	BII	BIII	BIV	BV	Square-Sum
1a		+·408	+·598	+·545	+·373	+·189	+·063	1·000
2a		+·577	+·676	+·309	—·105	—·267	—·178	1·000
3a		+·707	+·621	·000	—·258	·000	+·218	1·000
4a		+·816	+·478	—·218	—·074	+·189	—·126	1·000
5a		+·913	+·267	—·244	+·167	—·085	+·028	1·000
1b		+·913	—·267	—·244	—·167	—·085	—·028	1·000
..		..	..	..	..	..	..	..
Square-Sum (v)		5·000	3·000	1·000	·500	·300	·200	10·000

(b) *Factor-Measurements.* To obtain the partial regressions or 'weights' for the several categories ($W' = V^{-1}F'$) it is only necessary to divide each column of factor-saturations by the square-sum shown at the foot. We then apply these weights to the standardized form of the initial matrix of marks. The weighted sums yield the values for the several factor-measurements ($P = W'M$, where $M = D^{-\frac{1}{2}}M_1$). These are given in Table V.

As I have endeavoured to show in the Appendix, where the scale pattern is perfect, the factor-measurements can be obtained more directly by the formulæ used in calculating the constants for curvilinear regressions. These figures have already been tabulated. Thus if the reader will turn to the tables of orthogonal polynomials given by Fisher and Yates,¹ and take the values of ξ_i for $i = 1, 2, \dots, 5$, normalizing each column by the aid of the square-sums printed at the foot, he will find that *the figures so obtained are identical with the factor-measurements* given in Table V (except for the first factor which always has the value $+1/\sqrt{n}$). This set of tables may be used for a battery containing any number of items from 2 to 75.

Taking the factor-saturations as weights, it will be found that we can reconstruct the original standardized measurements *exactly* by the usual equation $M = FP$. With this method of factor analysis, therefore, it can no longer be objected that the initial data are not reproducible from the factorial results.

TABLE V. FACTOR-MEASUREMENTS FOR PERSONS (P)

Type of Person	5	4	3	2	1	0	Square-Sum
Factor G	+·408	+·408	+·408	+·408	+·408	+·408	1·000
" BI	+·598	+·359	+·120	—·120	—·359	—·598	1·000
" BII	+·545	—·109	—·436	—·436	—·109	+·545	1·000
" BIII	+·373	—·522	—·298	+·298	+·522	—·373	1·000
" BIV	+·189	—·567	+·378	+·378	—·567	+·189	1·000
" BV	+·063	—·315	+·630	—·630	+·315	—·063	1·000
Square-Sum	1·000	1·000	1·000	1·000	1·000	1·000	6·000

Scale Analysis. Dr. Guttman's 'component weights and scores' are derived directly from the initial table of data. He computes no coefficients of correlation or contingency, nor does he print any matrix of product-sums. The 'component weights' (if required) he calculates only *after* the 'component scores' have been obtained. His threefold theoretical approach has been succinctly summarized in the preceding paper (p. 1 above); and the algebraic equations thus obtained (given in his previous publications) bring to light a number of new and interesting features. Yet I am doubtful whether, for the ordinary student,

¹ (16), Table XXIII, p. 70. The analogy between this mode of approach and that adopted by Fisher in fitting curved regression lines suggests that an adaptation of his treatment might be used to test significance in the case of empirical data: but that is a question I hope to take up in a later paper.

the working procedure which the equations suggest would be easier to follow or quicker to apply than the ordinary factorial procedure, especially as that can be appreciably simplified in the case of an artificially perfect scale.

(a) *Component Scores.* The table given by Dr. Guttman as obtained by his own procedure is reproduced in Table VI. Like the factor-measurements given in Table V, the scores given by Dr. Guttman are orthogonal by factors (i.e., the rows are uncorrelated); but, unlike the former, they are not orthogonal by persons. His figures in fact are identical with those to be found in the table printed by Fisher and Yates ¹ ((16), p. 70).

TABLE VI. SCORES FOR PERSONS (GUTTMAN)

Component	Persons						Square-Sum
	5	4	3	2	1	0	
Constant	1	1	1	1	1	1	6
I	5	3	1	-1	-3	-5	70
II	5	-1	-4	-4	-1	5	84
III	5	-7	-4	4	7	-5	180
IV	1	-3	2	2	-3	1	28
V	1	-5	10	-10	5	-1	252

(b) *Component Weights.* Having dealt with the problem of "quantifying the rank order of people," Dr. Guttman turns to that of "quantifying the category values of items." These values he terms the 'component weights.' The figures that he gives ((18), Table 5, p. 329) are not proportional to my figures either for the 'factor-saturations' or for the 'regression weights.' They are, however, indirectly related. We can in fact derive the proportionate values for Dr. Guttman's 'weights' by dividing the saturations for each category (given in Table IV) by the square root of the total number of positive entries for that category, that is, multiplying them by $1/\sqrt{N}$. We then arrive at the figures shown in Table VII. The effect of this procedure is to allot the same weight for the general factor to every category in the list, while, with the remaining factors, the figures obtained are now simple multiples of the smallest figure in each column. In Dr. Guttman's table of 'principal component weights' the figures are either identical with these multiples or with simple fractions of them.

TABLE VII. UNSTANDARDIZED FACTOR-SATURATIONS (U)

Category	Multiplier	G	BI	BII	BIII	BIV	BV
1a	$1/\sqrt{1} = 1.000$	.408	.598	.545	.373	.189	.063
2a	$1/\sqrt{2} = .707$	.408	.478	.218	-.075	-.189	-.126
3a	$1/\sqrt{3} = .577$	.408	.359	.000	.149	.000	.126
4a	$1/\sqrt{4} = .500$	.408	.239	-.109	-.037	.094	-.063
5a	$1/\sqrt{5} = .447$	.408	.119	-.109	.075	.038	.012
1b	$1/\sqrt{5} = .447$	.408	-.119	-.109	-.075	-.038	-.012
..	..	..	..	..	..	..	..

He insists that his use of the conventional term 'weights' should not be taken to imply that such figures are needed to secure summational equations for computing the persons' scores. Nevertheless, as he remarks, "it so happens that the equations of internal consistency prove that there is a *reciprocal relation* between principal components for people and principal components for categories that does involve an additive (more precisely, an *averaging*) process, so we are justified in part in retaining the traditional nomenclature of 'scores' and 'weights.' " This is in effect a recognition of what I have elsewhere called the 'reciprocity principle.'²

¹ As Fisher and Yates point out, the figures are proportionate values only, and for convenience the figures given by the algebraic formulæ have been converted to whole numbers by the fractions shown at the foot of each table. As it happens, a table identical with Dr. Guttman's is also printed in the analysis of an example used by Mather to illustrate curvilinear regression (*Statistical Analysis in Biology*, 1942, p. 139). The student will find the detailed discussion in Mather's chapter exceedingly illuminating, particularly in regard to the analysis of variance.

² Cf. (10), pp. 487-94. In the notation there used, writing $F = LV\frac{1}{2}$, we have $M = LV\frac{1}{2}P$, where $LL' = I$ and $PP' = I$. Hence $P = V\frac{1}{2}L'M$ and $L = MP'V\frac{1}{2}$. Thus L can be regarded either as normalized factor-saturations (weights) for tests obtained by comultiplying tests or as normalized factor-measurements for tests obtained by comultiplying persons (in that case using $P'V\frac{1}{2}$ as weights). Similarly there will be two alternative interpretations for P .

Special Relations Arising with Perfect Scales. In discussing the theory of answer patterns, several writers have noted certain peculiar features characterizing the marginal totals of the columns and rows of ideal patterns like that illustrated in Table I. Definite relations are discernible both within and between (a) the series formed by the successive totals (weighted or unweighted) that express the 'difficulty' of the tests, (b) the series formed by the successive totals (weighted or unweighted) that represent the abilities of the persons or categories of persons, and (c) the series formed by the successive differences obtained by subtracting each total from the last. For example, as Walker has pointed out, if totals for persons form a linear scale, their first differences must be equal, and "the distribution of the raw scores (i.e., totals for persons) is equal to the first differences of the distribution of the number of persons passing each item correctly, so that the scores are completely determined by the difficulty values of the items": (see (6) and (11), pp. 53f.). Dr. Guttman's analysis generalizes these observations for all the components concerned.

The whole of his mathematical investigation of the subject ((18), pp. 334-61) is exceedingly instructive. In particular he shows that, if we examine the artificial type of answer pattern given by a 'perfect scale,' then, as he puts it in his paper (p. 2 above), we can discern "a definite law of formation for the entire set of principal components." Thus, with 'scalable' as with 'non-scalable' data, the principal components are of necessity uncorrelated; but, whereas with 'non-scalable' data they are in general completely independent, with 'scalable' configurations (as he points out) "all the components are perfect (curvilinear) functions of the rank order." In ordinary factor analysis (with linear regressions) instances in which a simple 'law of formation' may be found have often been noted in correlation tables of an artificial type (e.g., (10), pp. 300 and 306, and this *Journal*, II, Table V, p. 112); and with curvilinear factors we should expect the same structural relations to obtain as in the case of curvilinear regressions. The values for the latter can be built up from the differences¹ by the process used by Fisher ((8), p. 146, cf. (14), II, pp. 164f.)—a process not unlike the cumulative summation introduced by G. F. Hardy to facilitate the calculation of moments (cf. (4), 1941 ed., p. 442 and refs.). Thus I venture to suggest that the conditions imposed by the requirements of a perfect scale could perhaps be most succinctly expressed by the equations worked out for problems involving curvilinear regression, and that the same cumulative process may be used to facilitate the factorial analysis in dealing with data such as that considered above (see Appendix, pp. 21-3 below).

Hierarchical Classification. These numerical relations, however, are not likely to be accurately preserved in empirical tables obtained from actual tests or questionnaires. On the other hand, the distinctive sign pattern is a fairly constant feature except when there are marked irregularities in the nature of the items selected. On looking at the saturations in Tables III and VII, it is evident that the first factor (G) is a 'general' factor, specifying the 'genus' or 'universe' to which the items belong. The second (BI) is a bipolar factor distinguishing positive from negative item-categories. The third (BII) contrasts the easier tests with the harder; and so on. Similarly the factor-measurements in Table V classify and subclassify the persons according to successive levels of ability, i.e., according to the difficulty of the items which they can successfully answer. With a summational method of factorization such gradings can only be expressed by means of a progressive series of dichotomous classifications. Hence the automatic effect of the analysis is to classify both items and persons according to a somewhat artificial 'hierarchical scheme.'²

The curiously regular sign pattern which is thus produced is analogous to that obtained by applying the same method of factorization to a correlation table derived from specific factors only (cf. this *Journal*, II, p. 112, Table VC). However, as in that case (see *ibid.*, Table VB), so here, we

¹ The formulæ used are due originally to Tchebycheff (1): cf. Isserlis, L., 'A Note on Tchebycheff's Interpolation Formula,' *Biometrika*, XIX, 1927, pp. 87f., and (5) and (9). The difference equations reached by Dr. Guttman ((18) pp. 343-4) for expressing the relations between 'scores' (eqn. (38)) or between 'weights' (eqn. (46))—which (as he points out) reduce in the simplified case to a difference equation given by Tchebycheff—are among the most interesting of his results.

² As another writer has already pointed out (this *Journal*, V, p. 207), Ferguson has likewise adopted what is essentially the same view. Ferguson refers to my own explanation of bipolar factors as classificatory principles, and applies it to the factors obtained in analysing tests of varying difficulty. The factors so obtained, he says, "are exemplary of the principle of dichotomous classification": the first measures the average difficulty of the items; the second (Guttman's first or metric component) measures "the deviation of difficulty of the item from the average difficulty"; and similarly for all the rest ((12), p. 326). This, of course, is not quite the same as the simple types of divergent classification described which are termed 'hierarchical' by the older logicians. That requires a series of 'subdivided factors'; and these in turn introduce only $2^k - 1$ oscillations at each stage. The ordinary summational method yields cross-classifications as well as subdivisions; and this results in one more oscillation at each stage. (See this *Journal*, II, pp. 41-63, 'Subdivided Factors'.)

could, if we preferred, substitute Yule's triangular mode of factorization : this would show still more explicitly that each of the five bipolar factors dealt with five progressive levels of difficulty. I need not exemplify this point in detail here, since I have already illustrated it in a previous contribution.¹

IV. FACTOR ANALYSIS OF THE CORRESPONDING ANSWER PATTERN

Tabulation of Positive Categories Only. I now propose to consider the results that would have been obtained had we taken the experimental data embodied in Dr. Guttman's table and treated them according to the more usual routine methods. There will be two differences at the very outset. First, with two types of answer only, the actual performances would ordinarily be entered in a table specifying only 5 items, not 10 categories. Hence the rank of the product-sum matrix will be only 5, and we shall now obtain only 5 factors. Secondly, for purposes of correlation, the entries would be converted into deviation form. Since we have assumed $n = N - 1$, this reduction will make the number of degrees of freedom for persons the same as that for tests, and the effect of the change will be to weight both successful and unsuccessful answers according to the difficulty of the items. The resulting mark sheet is shown in Table VIII.

TABLE VIII. MARK SHEET FOR DR. GUTTMAN'S DATA IN UNITARY STANDARD MEASURE

Item	Persons						Square-Sum
	5	4	3	2	1	0	
1	5	-1	-1	-1	-1	-1	6 × 5
2	4	4	-2	-2	-2	-2	6 × 8
3	3	3	3	-3	-3	-3	6 × 9
4	2	2	2	2	-4	-4	6 × 8
5	1	1	1	1	1	-5	6 × 5
Total	15	9	3	-3	-9	-15	6 × 105

The 5×5 correlation table can be factorized by weighted summation in the usual way. The saturations and regressions thus obtained are set out in Tables IXA and IXB.

Factor-Saturations. The process of converting the observed test-marks to deviations about the mean has eliminated the former general factor (G), which was really a factor of difficulty for the several test-items ; and the new first factor now represents the general ability used in solving all such items. The factor-measurements for this factor therefore will measure the general ability of the persons tested.

The relation between the two sets of saturations will be clear if we compare the figures in Tables VII and IXA. Let us drop the first column in Table VII, and then restandardize the remaining figures so that the square-sum is still 1.000. This is equivalent to multiplying them by the values shown in the second column of Table IX (headed 'multiplier') ; and yields the results shown in the table. The sign pattern remains the same ; and the number of oscillations in the five successive rows is now 0, 1, 2, 3, and 4 respectively.

Factor-Measurements. To find the ordinary linear regression coefficients, each column of saturations is, as usual, divided by its square-sum. Then, on applying these coefficients (Table IXB) to the standardized test-measurements (obtained by dividing the marks in Table VIII by the square root of the square-sum), we obtain the factor-measurements for the five factors. It will be discovered that the figures thus computed are *precisely the same as those obtained by the previous procedure* (Table V above). They are, therefore, exactly proportional to Dr. Guttman's 'scores' ; and differ from his only by being expressed in unitary standard measure. Hence it may be fairly claimed that, so far as the essential results are concerned, the figures reached by an ordinary factorial analysis are identical with those reached by scale analysis.

¹ This *Journal*, IV, p. 76, Table IIB ; cf. also *Factors of the Mind*, p. 306, Table IIIB.

TABLE IX. FACTOR-SATURATIONS AND REGRESSIONS

Item	Multiplier	A. Saturations					B. Regressions				
		Factor					Factor				
		I	II	III	IV	V	I	II	III	IV	V
1	$\sqrt{6/5}$	·6547 +	·5976 +	·4083 +	·2073 +	·0690	·2182 +	·5976 +	·8166 +	·6908 +	·3451
2	$\sqrt{6/4}$	·8281 +	·3780 -	·1291 -	·3272 -	·2182	·2760 +	·3780 -	·2582 -	1·0906 -	1·0911
3	$\sqrt{6/3}$	·8783 +	·0000 -	·3651 +	·0000 +	·3086	·2928 +	·0000 -	·7302 +	·0000 +	1·5430
4	$\sqrt{6/2}$	·8281 -	·3780 -	·1291 +	·3273 -	·2182	·2760 -	·3780 -	·2582 +	1·0906 -	1·0911
5	$\sqrt{6/1}$	·6547 -	·5976 +	·4083 -	·2073 +	·0690	·2182 -	·5976 +	·8166 -	·6908 +	·3451
Square-Sum		3·0000	1·0000	·5000	·3000	·2000	·3333	1·0000	2·0000	3·3333	5·0000

Factor Analysis by Simple Averaging. In virtue of the symmetry that now characterizes each column of Table IX, the manner in which both saturations and regressions effect a hierarchical classification is here a little more obvious. Nevertheless, a set of 5 items and a sample of 6 persons do not lend themselves very neatly to classification by means of successive dichotomies. In order to exhibit the general principle in its most typical form we require an example with $n = 2^k$ items. Let us therefore drop the third item in Dr. Guttman's set of items. This will leave two persons in the same category, but may be all the more instructive on that account: (see top section of Table X).

The oldest device for securing what are essentially factor-measurements for the 'general factor' common to all the tests in a battery consists in just averaging the marks gained by each person. This is equivalent to multiplying the matrix of marks by a vector of weights which has every element positive and equal to $1/n$. When the number of items is an integral power of 2, the rest of the factor-measurements may similarly be obtained by multiplying the same matrix by vectors, each of which has every element numerically equal to $1/n$, but differs in sign pattern so that all the vectors have zero correlations with one another. If in every vector except the first the number of positive signs is equal to the number of negative, it will be possible to construct n such vectors for a battery of n items. On arranging them in order of increasing complexity, we shall find the demand for non-correlation introduces one additional oscillation in the sequence of plus and minus signs with each successive factor. The whole procedure is tantamount to taking as the first set of factor-measurements the simple averages of the marks; then averaging the residuals; then averaging the residuals from this second set of averages; and so on. It is in fact the method most commonly adopted for partitioning a given set of data in the analysis of variance.¹

To keep the square-sum for factors equal to that for items, it will be necessary to normalize the vectors by substituting $1/\sqrt{n}$ for $1/n$: see left hand panel of the middle section of Table X. On pre-multiplying the marks for the four remaining tests by this orthogonal matrix of weights we obtain the factor-measurements shown in the central panel. To recover the initial matrix of marks from the factor-measurements we have merely to take the transpose of the 'item weights,' and use it to pre-multiply the matrix of factor-measurements, as shown in the last panel of all.

Now let us compare the results thus obtained with the values in Tables IV, V, and IX (ignoring the item and the factor which have been suppressed). It will be seen that the 'item weights' and 'factor weights' in Table X have the same sign pattern as the regressions and saturations in the preceding tables, and that (except for the indeterminate sign of the zeros) the 'factor-measurements' also have the same sign pattern as the 'factor-measurements' in Table V, while their numerical values furnish a plausible approximation to the 'best fitting values' given by the more elaborate procedure.

¹ See Mather, *loc. cit. sup.*, pp. 61-4 (esp. sect. 23, 'The Principles of Partition'). Cf. also Burt, C., 'Factor Analysis and Analysis of Variance,' this *Journal*, I, 1947, pp. 5f. The factor-measurements thus computed are linearly independent, but the first is not statistically independent of all the others: nor do the figures give the closest possible fit. In order that the first condition may be fulfilled, we should weight each set of factor-measurements appropriately before subtracting them to obtain the successive residuals. For this modified method of obtaining factor-measurements direct from test-measurements, see Burt, C., 'The Influence of Differential Weighting,' this *Journal*, III, 1950, esp. pp. 115f.: it yields the same measurements as are obtained by simple summation (centroid formula with unreduced communalities). If the second condition is essential, the 'latent vectors' must be computed: these give the closest fit as judged by least squares.

TABLE X. FACTOR ANALYSIS BY SIMPLE AVERAGING

Items	Persons						Square-Sum	
	5	4	3	2	1	0		
1	5 -1 -1 -1 -1 -1						30	
2	4 4 -2 -2 -2 -2						48	
4	2 2 2 2 -4 -4						48	
5	1 1 1 1 1 -5						30	
I II III IV	Item Weights				Factor-Measurements			90 36 18 12
	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	6	3	0	
	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	3	0	-3	
	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	0	-3	0	
	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	1	-2	1	
1 2 4 5	Factor Weights				Initial Scores (Reconstructed)			30 48 48 30
	I	II	III	IV	5	-1	-1	
	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	4	4	-2	
	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	2	2	2	
	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	1	1	1	

The mode of classifying both items and persons implicitly adopted by summational methods of factorization will now be clearer. The first of the four factors—the general factor—has positive values for each of the ‘item weights’: when the values are equal, the effect is virtually to sum the correct responses; it thus specifies, in Dr. Guttman’s phrase, the ‘metric’ component. The second factor contrasts performance in the two harder items with performance in the two easier: in a certain sense, therefore, it is, as he says, an ‘intensity component.’ The third factor contrasts performance in the extreme items with performance in those of intermediate difficulty: it thus corresponds with Dr. Guttman’s component of ‘closure.’ If a name were needed for the fourth factor, I would call it an ‘alternating’ component; Dr. Guttman believes that it ‘portrays a new psychological concept’ which he terms ‘involution.’

Since the number of oscillations that characterizes each successive factor is primarily an outcome of the mathematical technique employed, it would seem better to envisage two ways of labelling the factors obtained. (1) First, it would be useful to have purely abstract designations merely indicating the *formal* properties of the several factors. And these designations would presumably differ according to our formal interpretation of the procedure. We may, for example, conceive the factors either as specifying the means (weighted or unweighted) of the observed measurements and of their successive residuals or as specifying the several types of independent comparison. (2) Secondly, we need concrete designations giving the *material* interpretation of the factors in this or that particular application; and the interpretation of course will vary according to the subject matter to which the analysis is applied.¹ From the formal standpoint (which alone concerns us here) the whole procedure, as I have suggested elsewhere, is rather like analysing the sounds of a musical instrument of complex timbre into a fundamental note with its successive overtones. The ideal test might be compared to a relatively pure note with little or no distorting noise; an analysis by subdivided factors is comparable to a Fourier analysis, where each additional harmonic is fainter than its predecessor and has twice the number of oscillations per second; and the recombination of factor-measurements by weighted summation is analogous to the compounding of simple harmonic motions at right angles to each other, as demonstrated (in the two-dimensional case) by Lissajous figures.²

¹ Spearman and Thurstone appear to use the word ‘factor’ exclusively in the latter sense. Elsewhere I have suggested that the term ‘component’ might be used for the purely mathematical ‘factor,’ and the name ‘factor’ reserved for the concrete interpretation.

² A still closer analogy is presented by the methods adopted for the ‘spectral analysis of aggregates’ in quantum physics; and, in view of the use of complex numbers in quantum theory, it is instructive to note that some of the problems arising in the factor analysis of frequency-data may also be solved by the use of components involving complex numbers (cf. Burt, C., ‘The Unit Hierarchy and its Properties,’ *Psychometrika*, III, 1938, pp. 151f.; also this *Journal*, IV, 1951, p. 20). In observing how the attempt to factorize a pattern of all-or-none responses gives rise to a set of oscillating components, it is impossible not to be reminded of the way de Broglie and Schrödinger have attempted to describe electrons in terms of oscillations in multidimensional space.

Conclusions. The foregoing examples will, I hope, sufficiently indicate that, with ideal tables of the type we have just been considering, a factor analysis can after all "show what a scale analysis will show," and that the use of the ordinary product-moment coefficient may lead to results which are by no means "misleading," even when the scale is perfect: a more general proof is outlined in the Appendix. But further it may, I think, also be claimed that a factorial approach such as I have described is particularly suitable with empirical data where the test-items do not form a perfect scale. This conclusion, it must be confessed, is based chiefly on experience in factorizing dichotomous test-items (like those composing such scales as the Binet-Simon or group-tests of the ordinary type) and qualitative assessments obtained for physical characteristics and for social conditions.

Now, as I emphasized at the outset, the problems discussed by Dr. Guttman originated in a rather different field. "The need for scale analysis," he tells us, "arises out of the fundamental problem of attitude scale and opinion polling of how to determine the dimensions of meaning which the questions asked have for the respondents: scale analysis affords a rigorous test for the existence of single-meaning for an area, and provides a rank order of individuals for such areas as are found scalable" ((18), p. 88). My own impression is that each of the various methods we have discussed has its special advantages and its special limitations, and the chief question awaiting solution is to determine still more precisely the particular conditions which give the preference to this method or to that. With this aim in view, it would, for instance, be instructive to study the results reached by applying the factorial methods described above to data obtained from questionnaires on attitudes or opinions, and by using Dr. Guttman's methods of scalogram analysis for data obtained with ordinary pass-or-fail test-items. Meanwhile, there can be no doubt that his intensive examination of the problems encountered in attempts to "quantify attributes" has brought to light many novel and instructive points, and should, in his phrase, go far to "stimulate rethinking about factor analysis in general."

V. SUMMARY

1. In analysing qualitative data, like those obtained from questionnaires or from tests where the answers are simply marked right or wrong, the numerical processes that arise are primarily those associated with counting rather than with measurement. In such cases, therefore, the appropriate statistical methods will be based on the principles worked out by Pearson and Yule for 'characters not quantitatively measurable' (to use Pearson's description) or 'attributes' (as Yule more succinctly termed them). Nevertheless, whether we are dealing with qualitative data or with quantitative, the problems that present themselves, and the procedures that are available, turn out in the main to be closely analogous. The study of scaling techniques, of total, partial, and multiple dependence, and of hypothetical factors or components, follows lines that are very much the same in either case.

2. As a preliminary to scaling the items that compose a questionnaire or test, the investigation of the unique type of 'answer pattern' that would be furnished by an ideal or perfect scale has proved highly illuminating, and has furnished various plausible indices for measuring the degree to which an empirical scale approximates to the pure or perfect type, e.g., the coefficients of 'hig' and 'unig' discussed by Walker, Ferguson, and others. The simplest of these has the same purpose and the same general character as the 'coefficient of reducibility' proposed by Dr. Guttman. Nevertheless, Dr. Guttman's more recent and intensive examination of the issues raised has elucidated much that was previously obscure, and revealed many new and fruitful problems.

3. Attempts to improve the construction and scaling of tests like the Binet-Simon series appear to indicate that for general purposes the most satisfactory method consists in a factorial analysis of the constituent items. It is applicable to empirical, as well as to ideal or perfect, scales. The results indicate a distinction between 'content factors' and 'difficulty factors,' and demonstrate a reciprocal relation between the difficulty of the test-items and the abilities of the persons tested.

4. For the study of qualitative data the most effective factorial technique would appear to be that originally proposed by Pearson for determining the orthogonal mean square regression lines (i.e., weighted summation with unreduced test-variances) or, as Pearson

himself preferred to call it, the method of 'principal axes.' Pearson's extension of regression to the determination of curvilinear regression lines suggests a generalization of factorial methods to deal with cases in which the relations are not linear but curvilinear. These modified methods are applicable both to measurement-data and to frequency-data. The use of regression equations of a degree higher than the first is analogous to the use of higher moments, and, with qualitative data, leads to the consideration of frequencies of higher order than the second. With ideal answer patterns the resulting factor-measurements (the latent vectors for persons) are proportional to the values tabulated by Fisher and Yates for orthogonal polynomials. It would seem therefore that Dr. Guttman's objections to the use of factor analysis with qualitative data may thus be fully met.

5. A factor analysis of Dr. Guttman's ideal 'scale pattern' (which includes negative as well as positive categories) yields figures for the factor-measurements which are identical with Dr. Guttman's 'component scores' when the latter are reduced to normalized form. The factor-saturations are related to his 'component weights,' but here the relation is less direct.

6. If an answer pattern of the more usual type (i.e., one which includes positive categories only) is substituted for the ideal 'scale pattern,' a factor analysis of the product-moment correlations still produces precisely the same factor-measurements. These factor-measurements exhibit a succession of dichotomous classifications, sub-classifications, and cross-classifications according to a hierarchical scheme, which is described by the sign pattern of the saturations and measurements. For practical work involving the analysis of empirical answer patterns certain abridged methods are available.

REFERENCES

1. Tchebycheff, P. L. (1859). 'Sur l'interpolation.' Reprinted in *Œuvres* (1907), I, 478-98, 541-60. St.-Petersbourg.
2. Pearson, K. (1901). 'On lines and planes of closest fit to a system of points in space.' *Phil. Mag.*, II, 6th Ser., 559-72.
3. Yule, G. U. (1910). *An Introduction to the Theory of Statistics*. London: Griffin.
4. Burt, C. (1921: 2nd ed. 1947). *Mental and Scholastic Tests*. London: Staples Press.
5. Allan, F. E. (1930). 'The general form of the orthogonal polynomials for simple series with proofs of their simple properties.' *Proc. Roy. Soc. Edin.*, L, 310-23.
6. Walker, D. A. (1931, 1936, 1940). 'Answer pattern and score scatter in tests and examinations.' *Brit. J. Psych.*, XXII, 73-86, XXVI, 301-8, XXX, 248-60.
7. Greenleaf, H. E. H. (1932). 'Curve approximation by functions analogous to the Hermite polynomials.' *Ann. Math. Stats.*, III, 204-17.
8. Aitken, A. C. (1933). 'On fitting polynomials to weighted data by least squares.' *Proc. Roy. Soc. Edin.*, LIV, 1-17.
9. Fisher, R. A. (1934). *Statistical Methods for Research Workers* (5th ed.). Edinburgh: Oliver & Boyd.
10. Burt, C. (1940). *The Factors of the Mind*. London: University of London Press.
11. Ferguson, G. A. (1941). *The Reliability of Mental Tests*. London: University of London Press.
12. Ferguson, G. A. (1941). 'The factorial interpretation of test difficulty.' *Psychometrika*, VI, 323-9.
13. Burt, C., and John E. (1942). 'A factor analysis of the Terman-Binet tests.' *Brit. J. Educ. Psychol.*, XII, 117-21, 156-61.
14. Kendall, M. G. (1946). *Advanced Theory of Statistics*. London: Griffin.
15. Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton: University Press.
16. Fisher, R. A., and Yates, F. (1948). *Statistical Tables*. Edinburgh: Oliver and Boyd.
17. Burt, C. (1950). 'The factorial analysis of qualitative data.' *Brit. J. Psych., Stat. Sect.*, III, 166-85.
18. Stouffer, S. A., et al. (1950). *Measurement and Prediction*. Princeton: University Press.
19. Burt, C. (1951). 'Test construction and the scaling of items.' *Brit. J. Psych., Stat. Sect.*, IV, 95-129.

APPENDIX

Let n be the number of items in the scale, and N the number of persons tested. According to Dr. Guttman, if the scale is perfect, the number of categories of persons between which it can discriminate will be $n + 1$. Let us therefore assume $N = n + 1$. Let j or k ($j, k = 1, 2, \dots, n$) be used to indicate the code-number assigned to the several items when arranged in order of difficulty, and a ($a = n, n - 1, \dots, 1, 0$) be the code-number assigned to the persons when arranged in order of ability: so that n denotes the easiest test and the ablest person. Finally, let x_{ak} stand for the mark obtained by the a th person in the k th item.

In a perfect scale, after the marks have been reduced to deviation form, there will be, for the k th item, k positive marks of $(n - k + 1)$, and $(n - k + 1)$ marks of $-k$. For each item the mean will therefore be $m_k = 0$; the square-sum (q_k say) will be $(n + 1)(n - k + 1)k$, and the variance (σ_k^2) will be $(n - k + 1)k$. Using the familiar formulæ for the sum of the first n numbers and for the sum of their squares, it is then easy to show that the sum of the n variances, $\sum \sigma_k^2$, will be $\frac{1}{6}n(n + 1)(n + 2)$, and that the total mark for the a th person will be $t_a = \sum x_{ak} = \frac{1}{2}(n + 1)(2a - n)$, and his mean mark therefore $\frac{1}{2n}(n + 1)(2a - n)$. The mean marks for the several categories of persons will consequently constitute a perfect linear scale diminishing by regular steps of $\frac{n + 1}{n}$; and their variance, σ_m^2 , will be $\frac{(n + 2)(n + 1)^2}{12n}$.

Let us now form a product-sum matrix by cross-multiplying the marks for every possible pair of items. The result will be, not a synclinal or hierarchical matrix, but what I have called an isoclinal¹ matrix. If we add each column of this matrix, and call the total for the k th item $\sum_j p_{jk}$,

we shall have² $\sum_j p_{jk} = -(n + 1) \sum_{j=0}^{n-k} t_j = \frac{1}{2}(n + 1)^2 k(n - k + 1) = \frac{1}{2}(n + 1)q_k = \frac{1}{2}(n + 1)^2 \sigma_k^2$.

For the correlations we have, by the usual product-moment formula, $r_{jk} = \frac{p_{jk}}{(n + 1)\sigma_j\sigma_k}$.

If we proceed to factorize the matrix of correlations by the method of weighted summation, then (as may readily be shown) the saturations for the first factor will be

$$f_k = \frac{1}{2}(n + 1) \frac{\sigma_k}{n\sigma_m} \quad (i)$$

We can verify this by applying the basic formula $f'_1 R = v_1 f'_1$ (where v_1 denotes the factor-variance and f'_1 the row-vector of factor-saturations used to weight the columns of the matrix). The factor-variance, v_1 , will be given by the sum of the squares of the n saturations, $\sum f_k^2 = \frac{(n + 1)^2 \sum \sigma_k^2}{4n^2 \sigma_m^2}$. On

substituting the values already reached for σ_k and σ_m , this expression reduces to $\frac{1}{2}(n + 1)$. Now, when we premultiply the j th element in the k th column of this matrix by the appropriate saturation of j , we obtain $f_j r_{jk} = \frac{1}{2}(n + 1) \frac{\sigma_j}{n\sigma_m} \cdot \frac{p_{jk}}{(n + 1)\sigma_j\sigma_k} = \frac{p_{jk}}{2\sigma_m\sigma_k}$. Summing all such values in the k th column, we have

$$\begin{aligned} \sum_j f_j r_{jk} &= \frac{1}{2n\sigma_m\sigma_k} \sum_j p_{jk} \\ &= \frac{1}{2n\sigma_m\sigma_k} \cdot \frac{1}{2}(n + 1)^2 \sigma_k^2 \\ &= \frac{1}{2}(n + 1) \cdot \left\{ \frac{1}{2}(n + 1) \frac{\sigma_k}{n\sigma_m} \right\} = v f_k. \end{aligned} \quad (ii)$$

Hence the values calculated for the saturations by eqn. (i) conform to the requirements of the saturations for the principal axes as stated by Pearson.

¹ See this *Journal*, IV, 1951, p. 47, and refs.

² The minus sign is due to the fact that, with Dr. Guttman's notation, the persons are numbered backwards.

Scale Analysis and Factor Analysis

The regressions (w_k), as usual, are obtained by dividing the corresponding saturations by σ_m . We thus have $w_k = \frac{1}{n} \cdot \frac{\sigma_k}{\sigma_m}$. The factor-measurement for the a th person will accordingly be obtained by taking a weighted sum of his standardized marks, viz.,

$$\begin{aligned} \sum w_k \frac{x_{ak}}{\sqrt{q_k}} &= \sum \left\{ \frac{\sigma_k}{n\sigma_m} \cdot \frac{x_{ak}}{\sigma_k \sqrt{(n+1)}} \right\} = \frac{\sum x_{ak}}{n\sqrt{(n+1)\sigma_m}} \\ &= \sqrt{\left\{ \frac{3}{n(n+1)(n+2)} \right\}} \cdot (2a-n). \end{aligned} \quad (\text{iii})$$

The sum of the squares of $(2a-n)$ for $a = 0, 1, 2, \dots, n$ will be $4\sum a^2 - 4n\sum a + (n+1)n^2 = \frac{1}{3}n(n+1)(n+2)$. Hence the effect of the expression under the radical is simply to reduce the values of $(2a-n)$ to standard measure; and these values will evidently form a perfect arithmetical progression decreasing by a constant difference of 2.

It follows that a factorial analysis applied to the standardized marks is fully capable of demonstrating whether or not the marks obtained by the several items form a perfect scale in Dr. Guttman's sense. To measure the degree to which any empirical set of marks approximates to a perfect scale, we could, if we wished, correlate the factor-measurements with the ranks of the several persons (or categories of persons). If we are not interested in determining the rest of the factors, it would for most purposes be sufficient to correlate the unweighted totals, without performing a complete factorial analysis to obtain the most appropriate least-squares weights.

Algebraic expressions for the remaining factors can be reached along similar lines. The order of the curve required to give a perfect fit is the same as the number of degrees of freedom available for the differences between the observed values, namely, $N-1 = n$. This will therefore be the order of the polynomial expression for the n th bipolar factor, i.e., for the last.

Let us reduce the rank order for the persons to deviation form, and write $x_{0i} = 1$ ($i = 1, 2, \dots, N$) for the general factor-measurements (which are the same for all persons), and $x_{1i} = a_i - \frac{1}{2}(N-1)$ for the first factor-measurements. Then the formulæ for all the factor-measurements can be expressed algebraically as follows.¹ I add in each case the numerical equations so given for the data examined in the preceding pages, where $N = 6$ (cf. Table V).

$$\begin{aligned} \text{G. } \frac{x_0}{\sigma_0} &= \frac{1}{\sigma_0} = \frac{1}{\sqrt{6}}. \\ \text{I. } \frac{x_1}{\sigma_1} &= \frac{1}{\sigma_1} x_1 = \frac{2}{\sqrt{70}} x_1. \\ \text{II. } \frac{x_2}{\sigma_2} &= \frac{1}{\sigma_2} \left\{ -\frac{1}{12} (N^2 - 1) + x_1^2 \right\} = \frac{\sqrt{3}}{4\sqrt{7}} \left\{ -\frac{35}{12} + x_1^2 \right\}. \\ \text{III. } \frac{x_3}{\sigma_3} &= \frac{1}{\sigma_3} \left\{ -\frac{1}{20} (3N^2 - 7) x_1 + x_1^3 \right\} = \frac{\sqrt{5}}{18} \left\{ -\frac{101}{20} x_1 + x_1^3 \right\}. \\ \text{IV. } \frac{x_4}{\sigma_4} &= \frac{1}{\sigma_4} \left\{ \frac{3}{560} (N^2 - 1) (N^2 - 9) - \frac{1}{14} (3N^2 - 13) x_1^2 + x_1^4 \right\} \\ &= \frac{\sqrt{7}}{24} \left\{ \frac{81}{16} - \frac{95}{14} x_1^2 + x_1^4 \right\}. \\ \text{V. } \frac{x_5}{\sigma_5} &= \frac{1}{\sigma_5} \left\{ \frac{1}{1008} (15N^4 - 230N^2 + 407) x_1 - \frac{5}{18} (N^2 - 7) x_1^3 + x_1^5 \right\} \\ &= \frac{\sqrt{7}}{20} \left\{ \frac{11567}{1008} x_1 - \frac{145}{18} x_1^3 + x_1^5 \right\}. \end{aligned}$$

The values for $\sigma_h = \sqrt{\sum x^2}$ ($a = 1, 2, \dots, N$), the constant used to reduce the values to unitary standard measure, can be obtained by the following equation:

$$\sum x_h^2 = \frac{(h!)^4}{(2h)!(2h+1)!} N(N^2-1)(N^2-2^2) \dots (N^2-h^2). \quad (\text{iv})$$

It will be seen that, although the five sets of factor-measurements are uncorrelated, *they can nevertheless all be expressed as polynomial functions of one factor only, namely, the first*. The order of the polynomial increases step by step; the even factors contain only even powers, while the odd contain

¹ The unstandardized values, denoted by x_h and given within the brackets, are proportional to, but not identical with, those given by Dr. Guttman, which give proportionate values in terms of the smallest integral numbers, like those tabulated by Fisher and Yates ((16), pp. 70f.). To obtain their tabulated figures, the above values for x_h must be multiplied by the fraction at the foot of their tables.

only odd powers : (hence the alternation between symmetry and reversal in the signs of the successive rows of Table V). For practical work the investigator may either take the values from the tables of orthogonal polynomials, or, if these are not available, build them up by means of the following relation ¹ :

$$x_{h+2} = x_1 x_{h+1} - b x_h, \text{ where } b = \frac{\sum x_{h+1}^2}{\sum x_h^2}.$$

For any factor the saturation for the j th item (f_j) can be obtained from the factor-measurements (p_i), by a process of cumulative summation, using the equation $f_j = \sqrt{\frac{(n+1)}{j(n-j+1)}} \times \sum p_i$, the summation extending to the first j measurements : e.g., for factor BI, taking the figures from line 2 of Table V, $f_2 = \sqrt{\frac{6}{2 \times 4}} \times (.598 + .359) = .866 \times .957 = .8281$, as given in line 2 of Table IXA.

The factor-variances are given by the equations :

$$v_1 = \frac{n+1}{1 \cdot 2}, v_2 = \frac{n+1}{2 \cdot 3}, \dots, v_r = \frac{n+1}{r(r+1)}, \dots, v_n = \frac{n+1}{n(n+1)} = \frac{1}{n}.$$

Their sum is

$$\begin{aligned} \sum v_h &= (n+1) \left\{ \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \dots + \frac{1}{n(n+1)} \right\} \\ &= (n+1) \left\{ \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \dots + \left(\frac{1}{n} - \frac{1}{n+1}\right) \right\} = (n+1) \left(1 - \frac{1}{n+1}\right) = n. \end{aligned}$$

¹ The relation is due to Tchebycheff (1). Cf. also (5) and Isserlis, L., 'A Note on Tchebycheff's Interpolation Formula,' *Biometrika*, XIX, 1927, pp. 87f.

COVARIANCE ANALYSIS AND ITS APPLICATIONS IN PSYCHOLOGICAL RESEARCH

By NEIL GOURLAY

Department of Education, University of Birmingham

I. Problem. II. Main Uses of Covariance Analysis. III. Comparison of the Precision of Covariance with that of other Procedures. IV. Summary and Conclusions.

I. PROBLEM

Fisher's analysis of variance is now a commonplace in the field of educational and psychological research : its principles and methods of application are expounded in all modern statistical text-books. But, when we turn to its correlate, the analysis of covariance, it is probably true to say that very few text-books provide an adequate treatment of the subject, although some give an excellent account of a particular aspect, and that none gives a complete and critical discussion of its uses in psychology. The wide variation in approach (compare, for example, Lindquist (1), Snedecor (2), McNemar (3), and Kendall (4)) is itself a source of confusion to all but systematic students of statistical method.

This paper therefore attempts to review the main uses of the analysis of covariance with special reference to its applications in psychological research. Its primary object will be to show that, although the statistical analysis remains more or less the same in all cases, the interpretations of the results must vary greatly with the conditions under which the data were obtained. Further, in indicating the correct interpretations for different experimental situations, the opportunity will be taken of pointing out some of the wrong interpretations made in applying the technique. A comparison will also be made of the precision of covariance with that of other methods ; and it will be shown that there are frequent occasions when a simpler technique can be employed with little loss in precision and a considerable reduction in the amount of computational labour. Bartlett (5)¹ has published an excellent but brief paper on covariance analysis. But this appeared in a journal which is not readily accessible to the average educational and psychological investigator, and was written mainly for the agricultural experimenter ; moreover, the misinterpretations of which Bartlett speaks are still to be found.

II. MAIN USES OF COVARIANCE ANALYSIS

Whatever other interpretations may be given, an analysis of covariance can always be regarded as a means of describing the manner in which any particular variate varies with respect to other variates within a given number of groups or populations. Covariance is therefore always applicable when samples from the same or different populations are measured for two or more variates, whether or not the measurements are made at the same time. However, every hypothesis put for-

¹ I have to acknowledge that, while most of my ideas on the analysis of covariance had taken shape before I read Bartlett's paper, their final form owes much to it. I am also indebted to him for certain comments he has been good enough to make on my manuscript and take this opportunity of expressing my thanks.

ward for testing requires that some set of statistical conditions shall be satisfied before the technique can be validly applied. These differ with the different types of hypothesis, and are not always clearly defined in the text-books. They will be stated as explicitly as possible in the course of the following discussion.

In this section the various applications will be considered with reference to three types of case :

- (a) where covariance is used to increase the precision of an experiment ;
- (b) where it is a means of analysing the relation between variates which have been measured after treatment of experimental groups ;
- (c) where it is applied to data obtained from non-random groups.

For simplicity, the discussion will be confined to the case of two variates only, i.e., we shall consider the variation of one variate, the dependent, relative only to one other, the independent, variate. This is by far the commonest situation in educational and psychological research ; and a similar argument holds when more than one independent variate is involved.

(a) *The Analysis of Covariance as a Means of Increasing the Precision of an Experiment.* This is probably the most important function that the analysis of covariance can fulfil in educational and psychological research. It is the only application considered by Fisher and Lindquist in their text-books ((6), (7), and (1)) : and it is one of the two discussed by Edwards (8). On the other hand, it receives scant treatment from Snedecor (2), and is not treated at all in many other text-books, e.g., Johnson's (9). Johnson admittedly, in one of his examples, speaks of covariance as 'increasing the precision of the experimental comparisons' ; but, as I shall point out later, the phrase is not legitimate in such a context.

When one speaks of increasing the precision of an experiment by the use of covariance analysis, it is evidently implied that this increase is relative to the precision obtained by other methods of analysis. Here we shall compare covariance with the simple analysis of variance only. For ease of exposition, let us make the comparison in terms of a simple experiment designed to compare the differential effects of, say, two or more methods of teaching a topic in some school subject. For both techniques the experimental procedure is the same except in one respect : in particular it is essential that the experimental groups shall form *random* samples from the given population. But, when using covariance, besides applying an appropriate criterion as a test at the conclusion of the experiment, it is also necessary that the groups shall be tested prior to the experiment and that the test adopted shall correlate well with the criterion. Often the same test is used, thus ensuring the maximum correlation. In the covariance analysis the scores in the initial and final testing become the values for the independent and dependent variate respectively. The two variates may be conveniently referred to as the *initial* and *final* variates.

The null hypothesis is the *same* for both analyses : otherwise it would be meaningless to talk of greater precision in the case of covariance. A simple statement of the null hypothesis is that no significant difference exists between the methods of teaching. Alternatively, we may say that the scores for the experimental groups in the criterion-test are random samples from the same population ; or, less generally, that the population means for the two groups working under different teaching methods are equal, i.e., $\mu_{py} = \mu_{qy}$, where μ_{py} and μ_{qy} are the population means for the *p*th and *q*th methods groups in the criterion-test (*y*).

An analysis of variance tests the hypothesis by providing two independent estimates of the common population variance—the 'between methods' and 'within methods' variances based on $(S - 1)$ and $(N - S)$ degrees of freedom respectively, where *S* is the number of methods and *N* is the total number of subjects employed in the experiment. The variances are tested for heterogeneity by means of Snedecor's *F*-test or Fisher's *z*-test. Necessary assumptions in making the test are that the scores 'within methods' are normally (or approximately normally) distributed and that the methods do not give rise to significant differences in the variances for the different experimental groups (the condition of homogeneity of variance).¹

¹ Although in this paper I have some criticism to make of Johnson's treatment of the analysis of covariance (9), his treatment of the analysis of variance is to be recommended in that he always tests for homogeneity of variance—a precaution that many investigators often overlook. But, rather inconsistently, he does not consider it necessary to test the underlying assumptions required for an analysis of covariance.

An analysis of covariance, on the other hand, utilizing the fact that the differences between final scores are to some extent a reflection of differences between the initial scores, removes the variance due to these initial differences from the final variances. It thus reduces the size of the error variance, and so increases the precision of the experiment. The algebraic details are given in the text-books (e.g., (1), pp. 191-3). Briefly the steps are :

1. An analysis of the sums of squares for the initial and final scores (i.e., the independent and the dependent variate respectively).
2. An analysis of the sum of products for the two variates.
3. Calculation of the residual sum of squares for the dependent variate (*a*) when the *total* sums of squares and products are used to fit a regression line to the total data and (*b*) when the *within methods* scores of squares and products are used to fit a regression line of common slope to each of the methods groups.
4. The subtraction of the residual sum of squares for *within methods* from the *total* residual sum of squares to give the *between methods* residual sum of squares.

One degree of freedom is lost in fitting the regression line ; and the results of the analysis can therefore be tabulated as follows :

Residual sums of squares					d.f.
Between methods	..	..	..	..	$S - 1$
Within methods	..	..	..	..	$N - S - 1$
Total	..	..	..	..	$N - 2$

On the null hypothesis the residual sums of squares *between methods* and *within methods* provide independent estimates of the same variance, namely, the residual variance in the parent population. Hence the hypothesis is still tested by the *F*-test. The underlying assumptions are clearly stated by Lindquist ((1), pp. 195, 196) :

- (i) The methods groups must be *random* samples from the *same* population;
- (ii) The regression of final on initial scores must be linear ;
- (iii) The regression coefficients must not differ significantly from group to group ;
- (iv) Within any group the dependent variate must be distributed normally for any given value of the independent variate, and have a constant variance, the same for all groups, namely, the residual variance already mentioned.

For each of these assumptions tests are available ; and, strictly speaking, they should form part of any analysis of covariance. The test for condition (iii) is given by Kendall (4), Snedecor (2), and Jackson (10). The test for homogeneity of residual variance is similar to that used in the analysis of variance : an example of its application in covariance analysis will be found in Jackson ((10), pp. 75-7). No assumption need be made about the distribution of the independent variate ; and consequently the condition of normality is not necessary (condition (iv) applies only to the partial variate).

What is the actual gain in precision achieved by covariance as compared with a simple analysis of variance ? Since precision is measured by the reciprocal of the error variance ((7), p. 182), the question can be answered by comparing the error variances. The ratio of the two precisions proves to be very nearly $(1 - r^2)$, where r is the product-moment correlation within groups. This follows from the well-known fact that the residual sum of squares within groups for the dependent variate is equal to $\sum y^2(1 - r^2)$, where $\sum y^2$ is the sum of squares within groups for that variate. Remembering that the residual sum of squares is based on $(N - S - 1)$ d.f. as compared with $(N - S)$ d.f. for the sum of squares 'within groups,' it will be seen that the exact value of the precision ratio is $(N - S)(1 - r^2)/(N - S - 1)$, i.e., slightly more than $(1 - r^2)$.

In more complex experiments, where one or more factors are controlled and not randomized, the error variance may be an interaction term ; but the above result still holds, except that r is now derived from the sums of squares and products for the interaction term. It is to be noted, however, that, when the interaction term is based on a small number of degrees of freedom, the one degree that is lost in analysis of covariance may considerably reduce the gain in precision.

(b) *The Use of Covariance to Study the Relation between Variates Measured after Treatment of Experimental Groups.* This is the application with which Snedecor (2) is mainly concerned ; it is briefly considered by Edwards (8).

Here we are again dealing with an experimental situation, i.e., one where principles of randomization, statistical control, etc., are deliberately introduced. But now both variates are measured after the application of the experimental treatments : that is, both may be affected by the treatments, and thus show significant differences between groups.

Since both are now 'final' variates, it would appear arbitrary which we choose as the independent variate ; but in practice the choice is determined by the question which has prompted the analysis. Four possibilities arise¹ :—

- (i) We may know on *a priori* grounds that there are no differences between treatments for the independent variate ;
- (ii) Significant differences between groups are found for the independent variate² ;
- (iii) We may know *a priori* that there are real differences between treatments for the independent variate, although a test of group differences in this variate does not prove significant ;
- (iv) We may have no *a priori* knowledge about the differences between treatments for the independent variate ; but, on testing them, the differences are found to be non-significant.

In case (i) the situation is the same as that discussed in subsection (a) : and in applying the analysis of covariance we are (as before) obtaining a more precise test. To illustrate from the educational field, one could carry out an experiment on method, e.g., in the teaching of Arithmetic, and at the end of the experiment give tests in both Arithmetic and Intelligence to the groups. Then, if one could say *a priori* that there was no difference between the groups in intelligence (and the data supported this), the scores for the Intelligence test could be used to improve precision in testing the differences between groups in the Arithmetic test. This is not the best procedure. As already stated, maximum precision is obtained by testing the groups initially with a test which has a high correlation with the final test. However, a situation might arise in which an investigator through ignorance failed to set an initial test, and a statistician called in later tried to remedy the position by setting a test at the end which correlated well with the final test, and for which there was known to be no difference between groups.

Turning to case (ii), namely, that in which there are significant differences between groups in the independent variate, we shall assume that there are also significant group differences in the dependent variate : otherwise there would be little point in using covariance. That is the situation always assumed in text-books concerned with this type of application. Since in this case the statistical details are straightforward and trouble arises only when less technical descriptions are attempted, it will be advantageous to consider the former first. The algebraic procedure is the same as for subsection (a) ; and the same variances are involved in the *F*-test. The necessary assumptions (p. 27) are also the same. But the hypothesis tested is different.

We can say that what we are testing is the hypothesis that the regression line (for the dependent variate on the independent variate) is the same line (though not necessarily of the same gradient) for all treatments ; or that for a given value of the independent variate the estimated value of the dependent variate is the same for all treatments ; or again that the means of the bivariate populations corresponding to the different groups are collinear, the line of collinearity being the common regression line. Algebraically, if μ_{px} , μ_{py} and μ_{qx} , μ_{qy} are the population means for any two treatments *p* and *q* (*x* and *y* representing the independent and dependent variates respectively), the hypothesis is that $(\mu_{py} - \mu_{qy})/(\mu_{px} - \mu_{qx}) = \beta$, where β is the coefficient of regression in the common population.

It must be emphasized that this is not the same as the hypothesis involved in subsection (a), where $\mu_{px} = \mu_{qx}$ by virtue of the design of the experiment, and the hypothesis to be tested is simply $\mu_{py} = \mu_{qy}$ —the same hypothesis as is tested by the simple analysis of variance.³ It follows from the definition of the term 'precision' that this application of covariance cannot be regarded as a case in which the purpose is to increase precision. If the expression 'increase of precision' is to have any meaning, the hypothesis tested must be the same when the precision technique is used as when it is not.⁴

¹ As in subsection (a), it will be convenient to speak in terms of a simple experiment on methods of treatment.

² If an analysis of variance is used to test differences between groups in the independent variate, the minimum assumptions for an analysis of covariance (p. 27) must be supplemented by the assumption of normality (or approximate normality) within groups for the independent variate. Both variates must then be normally distributed.

³ Mathematically the hypothesis of subsection (a) may be considered as a limiting case of the other.

⁴ I therefore disagree with Snedecor ((2), p. 323) when he regards 'increase of precision' as one of the features of this application. The remarkable thing is that later (pp. 334, 335) he appears to be quite clear about the distinction.

Let us now attempt a less technical description of the hypothesis stated above. Still keeping to statistical terms, we can say that we are seeking to discover whether the relation which the dependent variate bears to the independent variate is the same *between groups* as *within groups*; more loosely, whether the 'between group' differences in the dependent variate are "adequately accounted for" (as the text-books sometimes put it) by the 'between group' differences in the independent variate. I give two examples.

(1) Two groups of young children, backward mentally and educationally, are assigned at random to two different kinds of environment. Some years later they are tested for intelligence and educational attainment. To determine whether a difference between the groups in attainment is accounted for by their difference in intelligence, an analysis of covariance might be suitably used.

(2) In an experiment to compare methods of teaching composition, the experimental groups may be tested at the end of the experiment for ability in composition and in the mechanical aspects of English. Once again we might use covariance to discover whether group differences in the second variate 'explain' the group differences in the first.

There is yet another popular way of expressing the purpose in this case. We can speak of testing for differences between groups in the dependent variate after the latter have been 'corrected' for differences between groups in the independent variate, or, as Snedecor put it ((2), p. 320), after 'due allowance is made for' such differences. So long as such expressions are understood to mean no more than is contained in the previous statistical statements, they are permissible. There is one interpretation which must not be made. It must not be thought that, when covariance is applied to make "due allowance for" the independent variate, we are then testing the same hypothesis as would be involved in an experiment where it was arranged *experimentally* that there was no difference between groups in that variate.¹

Case (iii) is easily dismissed. The interpretation of an analysis of covariance for this situation is the same as for case (ii).

Case (iv) might at first sight be put with case (i).² But, if one considers the case in which normally distributed groups show no significant differences for either variate but a significant overall difference when a Δ -test is applied,³ it will be seen at once that the identification would be wrong. The significant overall difference must be due to real differences between groups in either the one variate or the other or in both; and the data do not allow us to accept one of these alternatives and reject the others. It follows that we cannot speak of using the one variate as the independent variate of a covariance analysis in order to increase the precision in testing differences as regards the other variate. Hence case (iv) cannot be identified with case (i), but should rather be interpreted along the lines adopted for case (ii).

(c) *The Analysis of Covariance Applied to Non-Random Groups.* In the previous applications the groups were random samples from the same population. In a third type of application they may be non-random or (to use Lindquist's expression) 'intact.' Instances from educational enquiries are the pupils in a school or grade or class. The variates may be measured at the same time or at different times (cf. Kendall (4), Johnson (9), McNemar (3), and Jackson (10)).

As before, we will consider the purely statistical aspect first. The algebraic procedure is the same: in this respect there are no differences between the three types. Further, the underlying assumptions are the same, except that the condition of randomness no longer holds. It must be emphasized that, with non-random groups, heterogeneity is more likely than in the previous cases. The researches quoted by Lindquist ((1), pp. 134-41, 223-7) suggest that, in dealing with intact school groups, heterogeneity of variance and correlation is the rule rather than the exception. The same is probably true of regression. Hence it is all the more urgent that the underlying assumptions of homogeneity of regression and residual variance should first be tested.

What statistical hypothesis is involved in this type? From subsections (a) and (b), the reader will see that the only valid hypothesis is that already stated for subsection (b), case (ii). If this is borne in mind, there will be no difficulty in examining the validity of

¹ An illustration of this point is a simple experiment on methods of instruction where the experimental groups do not receive the same total amount of instruction time. An analysis of covariance which allows for the differences in times will not necessarily give the same result (on applying the *F*-test) as an experiment where the times were made the same. The reader should be able to give the reason why this is so.

² Bartlett (5) does not appear to differentiate between cases (i) and (iv).

³ For an example of the use of the Δ -test, see this *Journal* (11).

less technical interpretations. Consider first whether covariance can be used to compare treatments, methods, etc., when applied to non-random groups instead of random samples. An educational investigator sometimes supposes that, if he uses initial scores (or intelligence-test scores) as the independent variate, he can validly apply covariance to non-random groups.¹ But, although the analysis makes an allowance for differences in initial scores or in intelligence-test scores (or in both), there is no guarantee that there may not be other differences between the groups operating to produce final differences which will be unattributable to the treatments. Unless other evidence is available, the danger of drawing a false conclusion is equally great when there are no differences between groups in the initial or final tests. Kendall's example ((4), pp. 240-2) illustrates this point. In his example the groups show no significant differences in the initial test but significant differences in the final test *after being subjected to the same treatment*.²

It follows that this third type of application, like case (ii), cannot be regarded as a method of increasing precision. I would therefore disagree with Johnson ((9), pp. 310-24) when he describes his use of covariance as 'increasing the precision of the experimental comparisons.' Obviously, since the same statistical hypothesis is involved as in subsection (b), case (ii), the interpretations which were suggested there can also be made for the present case. There is no need to repeat them. This does not exhaust the possibilities of this particular type. Kendall's example shows the value of covariance in prediction problems. In that example "a number of recruits are given a preliminary test to ascertain their suitability for a certain course of training; at the end of the training they undergo a proficiency test." One of the issues which the analysis is intended to decide is "whether the same rejection standards in the preliminary tests can be applied to all recruits, irrespective of their town of origin." The reader will note that, whatever specific purpose covariance is made to serve, the general purpose is always that of looking for homogeneity or heterogeneity—as in any analysis of variance. Jackson ((10), p. 74) expresses the same idea when he says that covariance decides whether the groups can be combined for the tests involved.³

III. COMPARISON OF THE PRECISION OF COVARIANCE WITH THAT OF OTHER PROCEDURES

In the preceding section I have endeavoured to indicate in what cases an analysis of covariance may be validly said to yield increased precision, and how the precision of covariance compares with that of a simple analysis of variance. I now propose to compare it with two other statistical procedures.

(a) *Matching Techniques.* Consider again a simple experiment on methods. It is well known that precision can be improved by matching the pupils, prior to the experiment, in one or more tests—in pairs if only two methods are to be compared, in threes if (as is somewhat rare) three methods are involved, and so on. The pupils in each matched group

¹ The first sentence of the second paragraph of McNemar's chapter on covariance (3) suggests that he takes this view.

² Kendall states (p. 243) that, by using the analysis of covariance to allow for initial differences, we obtain the effects of training alone on the groups. I would suggest that the residual differences between groups will be due partly to the effects of training on the groups, but partly to any difference in content between the preliminary and final tests that does not result from the training.

³ An unusual feature of Jackson's analysis ((10), p. 80) is that, in testing for differences between classes in the second test, after correcting for differences found in the first, he uses, not the variance derived from the residual sum of squares between classes (based on $(s - 1)$ d.f.), but the variance derived from the sum of squares due to departure from linear regression between classes (based on $(s - 2)$ d.f.). In other words, in correcting final differences for initial differences he is using not the regression *within classes* but the regression *between classes*. A valid interpretation can be given for Jackson's procedure (see Welch (12)); but I do not know whether he had this in mind.

are then assigned at random to one or other of the methods. For the final scores the analysis of variance will be tabulated as follows :

Sums of Squares				d.f.
Between methods	..	..	..	$p - 1$
Between matched groups	..	..	..	$n - 1$
Error	..	..	..	$(n - 1)(p - 1)$
Total	..	..	..	$np - 1$

where p is the number of methods and n the number of matched groups.

If the matching were perfect the error variance, it will be seen, would here furnish an estimate of the same variance as is estimated by the error variance in the analysis of covariance for the initial and final scores—namely, the residual variance within groups for the total population. As a rule matching is far from perfect, particularly if more than two methods are involved and the number of pupils is limited. Hence, on this count alone, the matching method never achieves as high a precision as the analysis of covariance. The precision of the matching procedure is also reduced owing to the fact that $(n - 1)$ d.f. are lost in 'taking out' the variance between matched groups. The loss amounts to 50 per cent. when only two methods are involved, 33 per cent. when three are involved, and so on. In some cases it may even render the method less precise than a simple analysis of variance. For this reason and because of the practical difficulty of matching pupils, experienced investigators seldom use a matching technique with experiments on methods and the like, if an analysis of covariance can be made.

Similar remarks apply to experiments on factorial design where other classifications besides that of treatments enter into the analysis. Thus, for an experiment on methods, replicated in several schools, the precision might perhaps be improved by matching the methods groups within schools ; but the same object can be achieved more easily and effectively by using random groups or even intact classes, and then applying an analysis of covariance.

(b) *Analysis of Variance of Gains (or Differences between Initial and Final Scores).* In his *Design of Experiments* (7), pp. 166-9) Fisher discusses certain alternatives to the analysis of covariance that have been used by investigators, namely :

- (i) an analysis of the variance of absolute gains of final over initial measurements ;
- (ii) an analysis of the variance of such gains expressed as a fraction or percentage of the initial measurements.

While recognizing that, in the usual agricultural experiment, such methods may not give results greatly different from an analysis of covariance, he points out that they entail arbitrary corrections for initial differences, and consequently never achieve the precision of a covariance analysis. Further, in the absence of other evidence, the experimenter has no assurance that the corrections are appropriate to his data, nor can he tell how much precision he has lost by employing them.

Fisher does not completely condemn these methods. And accordingly I propose to consider the first of them, and indicate (for experiments in the educational field) what conditions must be satisfied to obtain adequate results, i.e., results which do not fall short of the precision secured by an analysis of covariance.

The usual conditions of normality (or approximate normality) and homogeneity of variance will again hold good ; but in this case they need apply only to the gains (or differences). Let us denote the initial and final variates by X and Y , and consider the analysis of the variance of the differences $Y - X$. The hypothesis to be tested is obviously still the same as for an analysis of variance (of final scores) or covariance ; it will in fact be that $\mu_y - \mu_x$ (in the usual notation) is the same for all groups ; it thus reduces to the hypothesis that μ_y is the same for all groups (since μ_x is the same for all groups as a result of the experimental design).

Let σ_x^2 , σ_y^2 and ρ denote the population values of the variances within groups of the X and Y variates and of the product-moment correlation within groups respectively. Then the precision in the case of the covariance analysis is the reciprocal of the residual variance within groups, which, on the average, has the population value $\sigma_y^2(1 - \rho^2)$. For the precision in the case of the analysis of variance of differences we must consider the variance within groups which will have, on the average, the population value $\sigma_{y-x}^2 = \sigma_y^2 - 2\rho\sigma_x\sigma_y + \sigma_x^2$.

The loss in precision is therefore on the average ¹

$$\frac{1}{\sigma_y^2(1 - \rho^2)} - \frac{1}{\sigma_y^2 - 2\rho\sigma_x\sigma_y + \sigma_x^2} = \frac{(\sigma_x - \rho\sigma_y)^2}{\sigma_y^2(1 - \rho^2)(\sigma_y^2 - 2\rho\sigma_x\sigma_y + \sigma_x^2)};$$

or, expressed as a fraction of the maximum precision (given by the analysis of covariance),

$$\frac{(\sigma_x - \rho\sigma_y)^2}{\sigma_y^2 - 2\rho\sigma_x\sigma_y + \sigma_x^2} = \frac{\left(\frac{\sigma_x}{\sigma_y} - \rho\right)^2}{\left[\left(\frac{\sigma_x}{\sigma_y}\right)^2 - 2\rho\left(\frac{\sigma_x}{\sigma_y}\right) + 1\right]};$$

a quantity which is a function of ρ and the ratio σ_x/σ_y .

As will be seen, there is little loss of precision when σ_x/σ_y is approximately equal to ρ . Where σ_x/σ_y is not of the same order as ρ , and where a straightforward analysis of variance of differences would involve too great a loss in precision, the analysis of covariance can still be avoided by multiplying the values of the X and Y variates by factors, a and b say (so that $\text{var}[aX] \doteq \rho^2 \text{var}[bY]$), and then carrying out an analysis of variance of the differences $bY - aX$. This is sometimes a practical procedure.

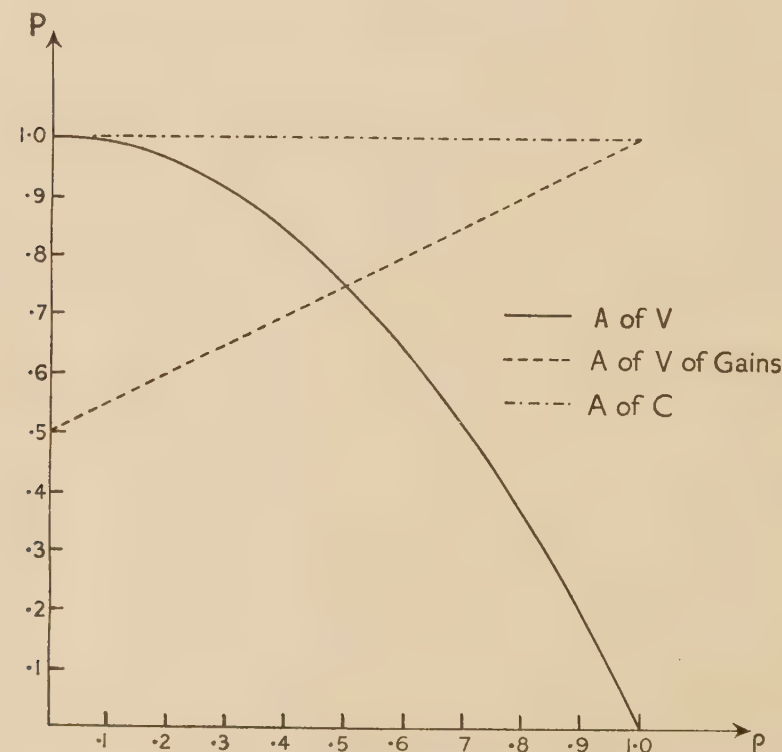


FIG. 1.—Precision of methods of analysis.

The diagram shows the variation of precision (P) with the correlation (ρ) for three methods of treating experimental data. The precision for the analysis of covariance has been taken as unity.

¹ To simplify the argument, I am here ignoring the loss of 1 d.f. incurred by an analysis of covariance.

In educational experimentation a common situation is to have σ_x equal (or approximately equal) to σ_y . This arises (a) when both the initial and final variate values are the scores on a standardized test, and (b) when scores on non-standardized tests have to be normalized before analysis in order that any pronounced skewness in score distribution may be removed and the condition of normality satisfied. Writing $\sigma_x = \sigma_y$ in the above expression, the loss in precision becomes $(1 - \rho)/2$. It will be seen that it is less than 10 per cent. if ρ is greater than .8 (which is often the case if the same test is used both at the beginning and end of the experiment) and only 15 per cent. for ρ equal to .7. Taking the precision of the analysis of covariance as unity, Fig. 1 shows how the precision of the analysis of variance and of an analysis of variance of differences varies with the value of ρ . The same considerations hold good in more elaborate experiments where several factors are controlled, the only difference being that σ_x^2 , σ_y^2 , and ρ now apply to the population values of the variances and correlation for the interaction term providing the estimate of error variance.

The reader may wonder why in such cases one should prefer an analysis of variance of differences to an analysis of covariance. The answer is simply that it saves considerable computational labour, often as much as two-thirds, particularly when the experiment involves several factors. It also avoids the loss of 1 d.f. involved in an analysis of covariance. A good practical procedure is to try the analysis of variance of differences first and to use the analysis of covariance in cases of borderline significance.

IV. SUMMARY AND CONCLUSIONS

1. The various uses of the analysis of covariance have been considered with special reference to its application in psychology and have been discussed under three main headings :

- (a) where covariance is used to increase the precision of an experiment ;
- (b) where it is a means of analysing the relation between variates which have been measured after treatment of experimental groups ;
- (c) where it is applied to data obtained from non-random groups.

2. In each case, special attention has been given to the following points ;

- (i) the necessary statistical conditions for the valid application of covariance analysis ;
- (ii) the exact nature of the statistical hypothesis involved ;
- (iii) the conclusions that can validly be drawn from the results of testing these hypotheses ;
- (iv) misinterpretations of covariance analysis which have been made in text-books and other publications.

3. An attempt has been made to define the particular conditions under which it is correct or incorrect to state that the use of covariance 'increases precision.'

4. The analysis of covariance has been compared in some detail with other methods of analysis—the analysis of variance for random and matched samples and the analysis of variance of gains (or differences). It has been shown that, while the maximum precision is normally obtained by means of covariance, in certain circumstances very little precision will be lost by applying an analysis of variance to the gains (or differences)—a technique which is computationally much simpler.

REFERENCES

1. Lindquist, E. F. (1940). *Statistical Analysis in Educational Research*. New York : Houghton Mifflin Co.
2. Snedecor, G. W. (1948). *Statistical Methods*. Iowa : Iowa State College Press.
3. McNemar, Q. (1949). *Psychological Statistics*. New York : John Wiley & Sons, Inc.
4. Kendall, M. G. (1945). *The Advanced Theory of Statistics, II*. London : Griffin and Co., Ltd.
5. Bartlett, M. S. (1936). 'A note on the analysis of covariance.' *J. Agr. Sci.*, XXVI, 488-91.
6. Fisher, R. A. (1946). *Statistical Methods for Research Workers*. London : Oliver and Boyd.
7. Fisher, R. A. (1949). *The Design of Experiments*. London : Oliver and Boyd.
8. Edwards, A. L. (1950). *Experimental Design in Psychological Research*. New York : Rinehart & Company, Inc.
9. Johnson, P. O. (1949). *Statistical Methods in Research*. New York : Prentice-Hall, Inc.
10. Jackson, R. W. B. (1940). *Application of the Analysis of Variance and Covariance Method to Educational Problems*. Bulletin XI, Dept. of Educ. Research, University of Toronto.
11. Rao, C. R., and Slater, P. (1949). 'Multivariate analysis applied to differences between neurotic groups.' *Brit. J. Psych., Stat. Sect.*, II, 17-29.
12. Welch, B. L. (1935). 'Some problems in the analysis of regression among k samples of two variables.' *Biometrika*, XXVII, 145-60.

THE DISTRIBUTION OF TOTAL RANK
VALUE FOR ONE PARTICULAR OBJECT
IN m RANKINGS OF n OBJECTS

By J. W. WHITFIELD
Psychological Laboratory, University College, London

I. Problem. II. Principle of Distribution. III. Moments. IV. Example.

I. PROBLEM

It is sometimes necessary to pick out for particular study one object only in a series, as when an experimental object is presented with control objects for preference ranking or some other form of judgement. A similar situation arises in experimental social psychology, when *one* person in a group is instructed to play a predetermined role, and each of the other members of the group is asked to rank his fellow members (including the experimental person) for certain characteristics. In these cases the experimenter is concerned with the ranks given to the specific object or individual. For this Kendall's (1948) treatment of the general m ranking problem is inappropriate ; and the following is suggested.

II. PRINCIPLE OF DISTRIBUTION

Consider one ranking of the n objects. The particular object can be assigned any rank from 1 to n ; and in the null case each value will have an equal probability. Interest lies in whether the object is ranked high or low, according to the experimental hypothesis under investigation. If a second ranking is considered, the distribution of the total of the ranks assigned to the object can take values from 2 to $2n$; but the null case no longer gives equal probability for each total value. The frequencies in the null case are made up in the following way. With a single ranking only ($m = 1$) we have :

	Rank Values										Total
	1	2	3	.	.	.	.	.	($n-1$)	n	
Frequency	1	1	1	.	.	.	.	.	1	1	n

Adding a second ranking to each possible value of the first, we obtain for $m = 2$:

1st Rank Value	Total Rank Values													Total (<i>f</i>)	
	2	3	4	.	.	.	<i>n</i>	(<i>n</i> + 1)	(<i>n</i> + 2)	.	.	(2 <i>n</i> – 2)	(2 <i>n</i> – 1)		(2 <i>n</i>)
1	1	1	1	.	.	.	1	1							<i>n</i>
2		1	1	.	.	.	1	1	1						<i>n</i>
3			1	.	.	.	1	1	1						<i>n</i>
.				.	.	.	.	.	.	.	.	.	.	.	.
<i>n</i> – 1							1	1	1	.	.	.	1	1	<i>n</i>
<i>n</i>								1	1	.	.	.	1	1	<i>n</i>
Total Frequency	1	2	3	.	.	.	(<i>n</i> – 1)	<i>n</i>	(<i>n</i> – 1)	.	.	.	3	2	1
															<i>n</i> ²

This illustrates the general principle that the frequency distribution for m rankings of n objects consists of the distribution for $m - 1$ rankings repeated n times in a simple staggered fashion—a principle which simplifies the computation of the exact values and facilitates the general expression for the second and fourth moments. In the appendix (pp. 38–40) tables of the exact probability values are given for m and n up to 8. For higher values of m or n a normal deviate approximation is possible.¹

III. MOMENTS

It will be seen that the distribution is symmetrical, and therefore the higher odd order moments are zero. The mean total rank value is $\frac{1}{2}m(n + 1)$, and deviations from this are to be tested. Let x be the deviation.

Variance

$$\begin{aligned}\Sigma x^2 &= n\Sigma x^2 + 2n^{m-1} \left\{ 1^2 + 2^2 + 3^2 + \dots + \left(\frac{n-1}{2} \right)^2 \right\} \dots \dots (n \text{ odd}) \\ \text{or} &= n\Sigma x^2 + 2n^{m-1} \left\{ \left(\frac{1}{2} \right)^2 + \left(\frac{3}{2} \right)^2 + \left(\frac{5}{2} \right)^2 + \dots + \left(\frac{n-1}{2} \right)^2 \right\} \dots (n \text{ even})\end{aligned}$$

(since Σx^2 is the sum of x^2 in the m ranking case, when all the n^m equiprobable occurrences are considered)

$$= n\Sigma x^2 + \frac{n^m(n^2 - 1)}{12}.$$

$$\text{Thus } n\Sigma x^2 = n^2\Sigma x^2 + \frac{n^m(n^2 - 1)}{12},$$

$$\text{and hence } \Sigma x^2 = \frac{mn^m(n^2 - 1)}{12},$$

and the variance therefore

$$= \frac{m(n^2 - 1)}{12}.$$

Fourth Moment

$$\begin{aligned}\Sigma x^4 &= n\Sigma x^4 + 2 \times 6\Sigma x^2 \left\{ 1^2 + 2^2 + 3^2 + \dots + \left(\frac{n-1}{2} \right)^2 \right\} + \\ &\quad 2n^{m-1} \left\{ 1^4 + 2^4 + 3^4 + \dots + \left(\frac{n-1}{2} \right)^4 \right\} \dots \dots (n \text{ odd}),\end{aligned}$$

¹ The results to be expected with perfectly random rankings may be compared with those which we should obtain if each of the m judges were to spin a perfect n -sided teetotum, with its sides numbered 1 to n , and we were then to add up the m scores: for the total scores obtained in this way the theoretical frequency-distribution can be deduced by the aid of the multinomial theorem. (I am indebted to Professor Burt for this observation. He points out that "the frequency for a total rank of x would then be given by the coefficient a^x in the product of the binomial expansions of $(1 - a^n)^m (1 - a)^{-m}$. With a little simplification the formula for the coefficients reduces to

$$\frac{m(m-1)\dots(x-1)}{(x-m)!} - m \cdot \frac{m(m-1)\dots(x-n-1)}{(x-n-m)!} + \frac{m(m-1)}{2!} \cdot \frac{m(m-1)\dots(x-2n-1)}{(x-2n-m)!} - \dots,$$

where the series is continued so long as none of the factors are negative. For tail-values, only the first one or two terms will require to be calculated. To find the probabilities we divide by n^m . The moments can be obtained in the usual way from the factorial moment generating function.")

This approach yields the same figures as are given in the appendix, and would make it possible to deal with the more general type of problem when equiprobability cannot be assumed.

with corresponding values for n even,

$$\begin{aligned} &= n \sum_{m-1} x^4 + 12 \sum_{m-1} x^2 \frac{n(n^2-1)}{24} + 2n^{m-1} \frac{n(n^2-1)(n^2-4)}{160} + 2n^{m-1} \frac{n(n^2-1)}{96}, \\ &= n \sum_{m-1} x^4 + (m-1)n^{m-1}(n^2-1) \frac{n(n^2-1)}{24} + \frac{n^m(n^2-1)(n^2-4)}{80} + \frac{n^m(n^2-1)}{48}. \end{aligned}$$

$$\text{Therefore } n \sum_{m-1} x^4 = n^2 \sum_{m-2} x^4 + (m-2)n^{m-1}(n^2-1) \frac{n(n^2-1)}{24} + \frac{n^m(n^2-1)(n^2-4)}{80} + \frac{n^m(n^2-1)}{48}.$$

$$\begin{aligned} \text{Hence } \sum_m x^4 &= \frac{n^m(n^2-1)^2}{24} \{ (m-1) + (m-2) + \dots + 1 \} + \\ &\quad \frac{mn^m(n^2-1)(n^2-4)}{80} + \frac{mn^m(n^2-1)}{48} \\ &= \frac{n^m(n^2-1)^2 m(m-1)}{48} + \frac{mn^m(n^2-1)(n^2-4)}{80} + \frac{mn^m(n^2-1)}{48}. \end{aligned}$$

$$\text{Therefore } \mu_4 = m(n^2-1) \left\{ \frac{(m-1)(n^2-1)}{48} + \frac{(n^2-4)}{80} + \frac{1}{48} \right\};$$

$$\begin{aligned} \beta_2 &= \frac{3(m-1)}{m} + \frac{9(n^2-4)}{5m(n^2-1)} + \frac{3}{m(n^2-1)} \\ &= 3 - \frac{6}{5m} - \frac{12}{5m(n^2-1)}, \end{aligned}$$

an expression which tends to the value 3 as m and (to a lesser extent) n increase. Since x is a discontinuous variable, a correction of $\frac{1}{2}$ is applied; and the normal deviate is therefore

$$\text{N.D.} = \frac{|\text{Total Rank Value} - \frac{1}{2}m(n+1)| - \frac{1}{2}}{\sqrt{\frac{m(n^2-1)}{12}}}.$$

The effect of tied rankings has not been considered. However, if ties are present, the true variance will be slightly less than that given above for the appropriate m and n . Thus the normal deviate obtained by this procedure will be slightly smaller than the true value, and the corresponding probability slightly too great. If a significant value is obtained when tested against the uncorrected variance, the presence of ties in the rankings will merely mean that the result has slightly more significance than the calculated figure would suggest.

IV. EXAMPLE

Eight judges rank seven objects. The experimental object has a total rank value of 43. What is the probability that this or a higher value would be obtained by chance?

$$\text{Variance} = \frac{8}{12} (49 - 1) = 32;$$

$$\text{Normal Deviate} = \frac{43 - \frac{1}{2} \times 8(7+1) - \frac{1}{2}}{\sqrt{32}} = 1.8556;$$

$$P = .0318 \text{ (approx.)}$$

To obtain the exact value from the tables we have to reverse the direction of the deviation, as the other tail only is tabulated. The corresponding deviation in the tables is therefore $m(n+1) - \text{Total Rank Value}$. In the example this is $64 - 43 = 21$. From the table for $n = 7$, $m = 8$ we find that a total rank value of 21 or less would be obtained by chance in .03113 of the trials. The approximation agrees closely with this value.

REFERENCE

Kendall, M. G. (1948). *Rank Correlation Methods*. London: Griffin and Co.

TABLES OF EXACT VALUES OF P FOR TOTAL RANK VALUES

The probability values given below are for deviations in the predicted direction only, and should be doubled if absolute probability values are required.

$m = 2$

Total Rank Value	n					
	3	4	5	6	7	8
2	.11111	.06250	.04000	.02778	.02041	.01563
3	.33333	.18750	.12000	.08333	.06122	.04688
4		.37500	.24000	.16667	.12245	.09375
5			.40000	.27778	.20408	.15625
6				.41667	.30612	.23438
7					.42857	.32813
8						.43750

$m = 3$

3	.03704	.01563	.00800	.00463	.00292	.00195
4	.14815	.06250	.03200	.01852	.01166	.00781
5	.37037	.15625	.08000	.04630	.02915	.01953
6		.31250	.16000	.09259	.05831	.03906
7		.50000	.28000	.16204	.10204	.06836
8			.42400	.25926	.16327	.10938
9				.37500	.24490	.16406
10				.50000	.34111	.23438
11					.44606	.31641
12						.40625
13						.50000

$m = 4$

4	.01235	.00391	.00160	.00077	.00042	.00024
5	.06173	.01953	.00800	.00386	.00208	.00122
6	.18518	.05859	.02400	.01157	.00625	.00366
7	.38272	.13672	.05400	.02701	.01458	.00854
8		.25781	.11200	.05401	.02915	.01709
9		.41406	.19520	.09722	.05248	.03076
10			.30400	.15895	.08746	.05127
11			.43200	.23920	.13578	.08057
12				.33565	.19783	.11987
13				.44367	.27280	.16968
14					.35860	.22974
15					.45190	.29907
16						.37598
17						.45801

TABLES OF EXACT VALUES OF P FOR TOTAL RANK VALUES—*continued*

The probability values given below are for deviations in the predicted direction only, and should be doubled if absolute probability values are required.

$m = 5$

Total Rank Value	n					
	3	4	5	6	7	8
5	·00412	·00098	·00032	·00013	·00006	·00003
6	·02469	·00586	·00192	·00077	·00036	·00018
7	·08462	·02051	·00672	·00270	·00125	·00064
8	·20988	·05469	·01792	·00720	·00333	·00171
9	·39506	·11730	·04032	·01620	·00750	·00385
10		·21680	·06916	·03241	·01499	·00769
11		·34863	·13824	·05877	·02749	·01410
12		·50000	·21984	·09799	·04683	·02417
13			·32224	·15201	·07479	·03912
14			·43904	·22145	·11287	·06018
15				·30517	·16202	·08844
16				·39969	·22241	·12476
17				·50000	·29321	·16962
18					·37234	·22304
19					·45683	·28436
20						·35229
21						·42493
22						·50000

$m = 6$

6	·00137	·00024	·00006	·00002	·0 ⁵ 8	·0 ⁵ 4
7	·00960	·00171	·00045	·00015	·00006	·00003
8	·03841	·00684	·00179	·00060	·00024	·00011
9	·10700	·02050	·00538	·00180	·00071	·00032
10	·23045	·04980	·01344	·00450	·00178	·00080
11	·40329	·10254	·02918	·00990	·00393	·00176
12		·18433	·05649	·01968	·00785	·00352
13		·29590	·09907	·03588	·01453	·00655
14		·42920	·15994	·06076	·02517	·01143
15			·23968	·09647	·04111	·01893
16			·33606	·14463	·06378	·02991
17			·44397	·20585	·09448	·04529
18				·27939	·13423	·06601
19				·36310	·18349	·09293
20				·45358	·24207	·12671
21					·30898	·16772
22					·38252	·21595
23					·46035	·27093
24						·33176
25						·39713
26						·46538

Distribution of Total Rank Value for One Particular Object in m Rankings of n Objects

TABLES OF EXACT VALUES OF P FOR TOTAL RANK VALUES—*continued*

The probability values given below are for deviations in the predicted direction only, and should be doubled if absolute probability values are required.

$m = 7$

Total Rank Value	n					
	3	4	5	6	7	8
7	·00046	·00006	·00001	·0 ⁵ 4	·0 ⁵ 1	·0 ⁵ 5
8	·00366	·00049	·00010	·00003	·00001	·0 ⁵ 4
9	·01646	·00220	·00046	·00013	·00004	·00002
10	·05167	·00732	·00154	·00043	·00015	·00006
11	·12986	·01971	·00422	·00118	·00040	·00016
12	·24691	·04492	·01005	·00283	·00096	·00038
13	·41015	·08936	·02125	·00610	·00208	·00082
14		·15820	·04070	·01206	·00416	·00164
15		·25305	·07162	·02209	·00775	·00307
16		·37012	·11686	·03787	·01359	·00543
17		·50000	·17824	·06122	·02260	·00915
18			·25574	·09388	·03584	·01477
19			·34714	·13716	·05445	·02293
20			·44794	·19170	·07954	·03432
21				·25717	·11205	·04972
22				·33216	·15259	·06987
23					·20137	·09543
24					·25802	·12693
25						·16466
26					·32161	·20864
27					·39065	·25856
28					·46315	·31377
29						·37329
30						·43586
31						·50000

$m = 8$

8	·00015	·00002	·0 ⁵ 3	·0 ⁶ 6	·0 ⁶ 2	·0 ⁷ 6
9	·00137	·00014	·00002	·0 ⁵ 5	·0 ⁵ 2	·0 ⁵ 5
10	·00686	·00069	·00012	·00003	·0 ⁵ 8	·0 ⁵ 3
11	·02393	·00252	·00042	·00010	·00003	·00001
12	·06447	·00743	·00127	·00029	·00009	·00003
13	·14129	·01854	·00327	·00077	·00022	·00008
14	·26078	·04033	·00750	·00178	·00052	·00018
15	·41564	·07805	·01555	·00379	·00111	·00038
16		·13638	·02959	·00745	·00222	·00077
17		·21768	·05210	·01369	·00415	·00144
18		·32034	·08573	·02369	·00736	·00259
19		·43826	·13263	·03887	·01242	·00443
20			·19392	·06071	·02007	·00727
21			·26918	·09065	·03113	·01152
22			·35622	·12983	·04654	·01763
23			·45115	·17888	·06724	·02616
24				·23771	·09406	·03770
25				·30540	·12769	·05289
26				·38017	·16852	·07233
27				·45953	·21655	·09656
28					·27135	·12602
29					·33203	·16095
30					·39728	·20139
31					·46543	·24714
32						·29772
33						·35237
34						·41012
35						·46982

NOTES AND CORRESPONDENCE

ON MAKING STATISTICAL ASSUMPTIONS

By A. S. C. EHRENBERG

Institute of Psychiatry, Maudsley Hospital

In psychological circles there are several schools of thought which regard mathematical and statistical methods, particularly if at all complex, with distrust. The criticisms appear usually to run along logical as opposed to empirical lines. Zangwill (1), for example, bases his criticisms on Chambers' pronouncement that 'it is at least very doubtful whether the concept of measurable quantity may be applied at all to psychological qualities' (2); and we are repeatedly told that the question of whether or not psychological qualities are measurable is a matter of logical or formal 'assumption.' At the same time the commonest line of attack on numerical methods in psychology is the complaint that their use involves too many 'assumptions.' Clearly, therefore, the nature of such so-called 'assumptions' and the possibility of error either in overlooking or in relying on them requires some examination, even without explicitly entering into the wider issues involved.

At least three kinds of error can occur in setting out the assumptions on which an argument is based. First, some of the assumptions implicit in the argument may be omitted or insufficiently explained. Secondly, the nature of the relevant assumptions and of their logical status may be misinterpreted. Thirdly, so-called assumptions may be introduced which cannot strictly be regarded as assumptions at all.

To illustrate the need for a closer scrutiny, I shall refer more particularly to the recent volume by Professor Harold Gulliksen on *Theory of Mental Tests* (3), and the comments passed on it in a review published in the *British Journal of Psychology* (*General Section*, XLIII, pp. 80-1) over the initials E. G. C.¹ Here the reviewer marshalled a set of 'assumptions' requisite for the application of statistical procedures in a manner which is admirably concise and so lends itself readily to a brief examination. In the review printed in this *Journal* (V, pp. 134-5) it was stated that in Gulliksen's book "the assumptions and definitions are carefully formulated": and this opinion also, as will be seen from what is said below, seems much too generous.

1. Failure to give *all* the assumptions inherent in an argument is a well-known fault; and failure to stress and explain the most important assumptions is still commoner. This type of error can be most simply illustrated by a familiar problem, namely, the use of normal distributions. Thus in Chapter XIV of Professor Gulliksen's book we find that the requirement of normality for Wilks' tests is not mentioned at all. More serious is the fact that Gulliksen emphasizes that almost all his formulæ are true, irrespective of any assumptions about the distributions involved. But the touchstone in applied mathematics is not so much whether the formulæ are 'true' (that the psychologist may leave to the mathematician), but whether they can in fact be legitimately applied. Now, unless we assume that the variables discussed by Gulliksen—'true' scores, 'error' scores, etc.—are in fact normally distributed, it is impossible to make precise and valid numerical deductions about the behaviour of our data from the formulæ set out by him. Normality is an inherent assumption in almost all correlational work (particularly that arising out of the theory of mental tests); but it is an assumption which must be fulfilled, not so much to make the formulæ true as to make them useful. Admittedly Gulliksen briefly mentions (3, p. 17) a need for assuming normally distributed *errors*, but he makes no reference to the distribution of the *raw* data. This rough and ready way of dealing with the distributional problem appears particularly unfortunate, since in practice many observations are not normally distributed. Indeed, the more advanced investigator often seems to do his level best to make the *joint* distribution of his data *not* normal by adopting systematic experimental designs (e.g., in working with an 'experimental' and a 'control' group).

At times there seems to be a peculiar difficulty in detecting the more isolated omissions. When, for instance, in the two reviews mentioned above, it is stated that the assumptions in Gulliksen's book "are scrupulously pointed out whenever they are made" or that they have been "carefully formulated," the reviewers may have been misled by the fact that *certain* assumptions have been stressed into supposing that *all* assumptions have been mentioned.

2. When assumptions are explicitly stated, there is often some confusion in the way their nature or their status is treated. Not infrequently it seems to be implied that they have to be accepted by an act of faith. In the review already quoted, for example, certain statistical assumptions are specified, and the writer then asks: 'How many psychologists would be *willing* to make them?'

¹ The editor of the *General Section*, to whom my comments were originally submitted, suggested that "it might be better to state the arguments more fully in the *Statistical Section*."

Once again examples can be found in the discussions of normality. Thus, in regard to the so-called 'error' scores, the reviewer tells us that 'without this or some similar assumption further progress cannot be made.'¹ But this normality is not an 'assumption' that the practical worker can make it merely by being 'willing' to do so. What is needed is an empirical confirmation: e.g., the data should be first fitted by some mathematical distribution function, possibly normal, possibly not. Such a procedure, however, turns the issue into a matter of descriptive statistics, not of simple assumption. If the data cannot be adequately described by normal distributions, then theoretical formulæ which assume normality should not be used.

Again, it is asserted that tests are sometimes 'assumed' to be parallel, i.e., to have equal means and equal standard deviations. In such cases it should be the preliminary duty of the investigator to determine whether the tests are 'parallel' (in this sense) or not. As before, the notion of being 'willing' just to *assume* such parallelism really indicates a serious omission in the analysis.

It seems possible that this mistaken attitude about 'having to make assumptions' may be a symptom of quite a different difficulty that is never explicitly stated. The reviewer's anxiety about assuming parallelism, for instance, might be due to his concern over the fact that in psychometric theory there seems to be no sound procedure for estimating the reliability of non-parallel data. If this is so, it indicates that a criticism, which at first sight might appear to be mere semantic quibbling, is needed to direct attention to points of obvious practical importance.

3. Perhaps the most dangerous examples of so-called 'assumption' are to be found in those statements which either cannot be verified or else are true by definition. When a writer claims to base an argument on certain assumptions, he really implies that his conclusions are capable of proof; whereas in the cases to which I am referring either they cannot be proved to be true (thus apparently requiring a 'willingness' to make the requisite 'assumption') or else they follow directly from the definitions and consequently are not in need of proof. It is perhaps instructive that much of the inconclusive work in psychological statistics—notably in arguments where a mathematical model is set up—may be traced to a confusion of this kind. As an instance, may I refer once again to the review in the *General Section*, where the writer refers to "the basic assumption that a person's raw score . . . is composed of a 'true' score and an 'error' score"? As it stands, it would be impossible to show that this statement is either true or untrue. We might perhaps amend it to read: "Let us try to describe the data by a model involving two variables, which may be called 'true' and 'error' scores, and may be defined as follows . . ." But the specification of a model is in itself not a matter of assumption or proof. For much the same reasons it is misleading to say that the 'true' and 'error' scores are *assumed* to be uncorrelated: such a statement would in effect be part of the definition of 'true' and 'error' scores in the model proposed.

The change in formulation is clearly of practical consequence. Instead of asking whether the mathematical model proposed is 'true' (in some way not explicitly defined), and assuming its 'truth' for the sake of the subsequent deductions, it becomes the investigator's business first to ascertain whether the model set up is internally consistent, and then whether it furnishes an appropriate scheme for describing the data.

Finally, three general points may be briefly considered. First, how is the concept 'assumptions' to be defined? Secondly, what is the effect of the foregoing arguments on the reviewer's evaluation of Professor Gulliksen's book? And lastly, how far is it legitimate to generalize the arguments here illustrated?

To attempt a formal definition of what is meant by the term 'assumption' is to invite criticism from various quarters. Yet some such definition seems needed, at least for *statistical* assumptions. We may begin by excluding broad and general axioms or postulates about the nature of empirical reality or about the bases of our logical systems: disagreement about these, though interesting and suggestive, rarely affects what the practical scientist does. Secondly, we may exclude scientific hypotheses as such, that is to say, undemonstrated propositions which are only provisionally put forward with a view to testing their plausibility in the light of the data. I suggest, therefore, that an 'assumption' may be defined as a proposition which may or may not be true in fact, but which must hold good if a given theoretical result is to be validly deduced. An obvious instance arises in the derivation of some formula incidentally required for estimating a parameter or testing significance. Some of the commonest statistical assumptions of this type assert the normality, linearity, or homoscedasticity of the distributions to which the formulæ are to be applied. In empirical science all such assumptions *must* in principle be verifiable; for, if one cannot discover whether some such incidental assumption is plausible in any given case, one cannot judge the plausibility of the final inference.

The verdict on Professor Gulliksen's book, pronounced by the reviewer in the *General Section*, declares that 'psychological research workers will . . . turn aside with feelings akin to horror.' Since it is stated that 'the algebra is worked out clearly enough,' the reviewer's reasons must presumably arise from his doubts over the assumptions made. Now, although on this latter count his argument

¹ He appears to be in error when he describes Gulliksen as assuming that the *standard deviation* of the 'error' scores is normally distributed (cf. (3), p. 17).

seems unacceptable, nevertheless his rejection of the book might still be endorsed for quite the opposite reason, namely, that the algebra and the statistical concepts appear, on closer scrutiny, to be decidedly confused.

Lastly, if it be agreed that the contentions here advanced are sound so far as they go, it is tempting to suggest that the main points may reasonably be generalized. Certainly, it seems likely enough that the widespread distrust of statistical methods in psychology is really due, not to the profounder logical or formal difficulties that are commonly said to beset all attempts at numerical measurement, particularly as these difficulties are seldom explicitly analysed; rather it would appear that it chiefly springs, however indirectly, from difficulties of a purely technical or empirical nature. Many psychometric techniques, for instance, prove to be extremely restricted in their applications, and in practice these restrictions are constantly forgotten. Others turn out to be inherently faulty as judged from a technical standpoint.

REFERENCES

1. Zangwill, O. L. (1950). *An Introduction to Modern Psychology*. London: Methuen & Co.
2. Chambers, E. G. (1943). 'Statistics in Psychology and the Limitations of the Test Method,' *Brit. J. Psychol. Gen. Sect.*, XXXIII, 191.
3. Gulliksen, H. (1950). *Theory of Mental Tests*. New York: Wiley.

ON MAKING STATISTICAL ASSUMPTIONS: A REPLY

Mr. Ehrenberg's *Note* on making statistical assumptions, which might perhaps more aptly be described as a review of a review, calls for a few comments. This is perhaps scarcely the place to enter on a controversy about assumptions. But, as the pilloried reviewer, I may say that considerations of space allowed me only 200 words in the *General Section* of the *British Journal of Psychology*, in which to describe Professor Gulliksen's book and express an opinion on its value to psychologists. Mr. Ehrenberg uses more than six times as much space in the present issue of the *Statistical Journal*, and apparently comes to the same conclusion—that the book is not to be recommended. Anyone who, distrusting my review, reads Gulliksen's book for himself would very soon discover the ambiguities in exposition (such as the one indicated in Mr. Ehrenberg's second footnote).

I believe that psychologists understood me when I used the word 'assumption' in the dictionary sense of 'something taken for granted.' Mr. Ehrenberg's claim that "one must always be able either to verify or disprove" assumptions does not find a place in the definitions of the *Oxford English Dictionary*; and it seems that he has forgotten that statistical methods are based on the principles of probability ('the calculus of uncertain inference') and that nothing can be proved or disproved by statistics.

E. G. C.

THE SCIENTIFIC STUDY OF PERSONALITY

A Note on the Review

By H. J. EYSENCK

Institute of Psychiatry, Maudsley Hospital

Reviewers have a great deal of latitude in the way in which they express their opinions ; and no author should be too sensitive when he finds his most dearly cherished thoughts given short shrift. But criticism must be of what the author has actually written ; it must avoid misquotations, factual errors, and other objectively fallacious arguments. The review of my book *The Scientific Study of Personality* (see this *Journal*, V, pp. 208f.) deviates so far from what is permissible that a note of protest appears necessary. I shall take up the reviewer's errors in numbered series, as they form an almost perfect example of the art of distortion.

1. *Change of sense through direct misquotation.* On p. 211 he prints the following sentence : "Burt's system of classification," he (i.e., Eysenck) says, "is some thirty years out of date, and discredited in contemporary psychology for many valid reasons." What I actually say (p. 41) is : "Burt's work is based on the rating of the strength of McDougallian 'instincts' in children by teachers. The choice of a system of classification some thirty years out of date . . ." No one reading the passage in its context can doubt that it is McDougall's system which I consider out of date ; the quotation is subtly changed in such a way as to give a completely erroneous impression.

2. *Change of meaning through indirect misquotation.* In the same paragraph, the reviewer, referring to Burt's 1915 *Report*, says that "Burt tells us, the correlations were based on 'items of behaviour systematically recorded.'" This is incorrect. Burt does not tell us at all what his correlations were based on. He merely discusses different possible ways in which ratings could be obtained, and which of these methods he considers superior.

3. *Change of meaning through incomplete quotation.* The reviewer quotes my sentence "Burt . . . has contrasted the factor of 'emotionality' with the factor of 'neuroticism'" (p. 115). He goes on to say that this is incorrect, and that Burt in fact considers the two to be "virtually identical." However, I continue my paragraph by quoting Burt's very words, bringing out clearly in what way he contrasts the two factors. Had the reviewer not broken off the quotation where he did the reader could have judged for himself the accuracy of my interpretation.

4. *Disparagement through imputation.* He discusses my claim that objective tests are superior to other types of tests. Because I do not quote the recent Symposium in the *Brit. J. Educ. Psychol.*, he concludes that I am "apparently unaware of the fairly comprehensive survey of the reliability and validity of different methods there brought together." He does not appear to consider the possibility that other causes than ignorance may be responsible ; and that other psychologists might not agree that the survey was "fairly comprehensive." It was not quoted because it was in my view so one-sided in its presentation, and so restricted in its coverage, as to be practically valueless for my purpose. To construe this into ignorance on my part of its very existence is unworthy of a scientific review.

5. *Disparagement through suppression and indirection.* The reviewer criticizes my use of objective tests by saying that the usual objection to such tests is that their reliability is low, and that their validity cannot possibly exceed their reliability. He complains that my tables seldom include reliability coefficients, and that those I give are by no means high. The tests specifically mentioned by him—dexterity, persistence, tapping, body-sway—are reported in the literature to have reliabilities which are quite satisfactory ; body-sway suggestibility test-retest reliability, for instance, is above .9, tapping is above .8, dexterity is between .7 and .9, while persistence has an estimated minimum reliability of .7. These values are not particularly low ; they would almost certainly be higher than reliabilities of the methods advocated by Burt and the reviewer. Surely the main point is that I have laid stress throughout my book on validity, rather than on reliability, which is merely a tool for the purpose of reaching higher validity. He is right in saying that an unreliable test cannot be valid. Surely the conclusion is that as these tests have been shown to be valid, they must *eo ipso* be reliable ? He suppresses the well-known fact that these tests have high reliabilities, and uses indirection to suggest that because they are believed (by him) to be unreliable, therefore they cannot be valid—suppressing the fact that they have been shown to be valid, and must therefore be reliable.

6. *Misrepresentation through falsification of argument.* On p. 210, he deals with my treatment of the problem of whether normal and pathological phenomena are continuous or not. He says : "Dr. Eysenck . . . prefers to assume that" they are continuous, and "having decided to assume continuity" goes on from there. The reader of the review would never guess from this that far from assuming continuity the main purpose of the experiments described in the book was to decide this question of continuity, and that I made no assumption at all about what I would find. I believe that the main contribution I have made on the experimental side to the study of psychiatry has been precisely this attempt to find conclusive evidence on this point as between the qualitative point of view and the quantitative one. It is possible to argue that I failed in this attempt ; my critic prefers to ignore it altogether, and to say calmly that I "assume" continuity. This is a complete misrepresentation of the situation.

7. *Misrepresentation through invention.* On p. 209 he says that in order to secure a unique and

invariant position I carry out "a rough graphical rotation." In fact the rotation is precise and algebraic. I do not quite see how one could carry out a graphical rotation under the special circumstances even if one wanted to do this. The "rough graphical rotation," therefore, is entirely an invention of the reviewer's: it finds no counterpart in my book.

8. *Misrepresentation through ignorance.* In dealing with criterion analysis, he says: (having quoted my aim in rotating as that of reaching a position for the factor-axis where the correlation between factor and criterion is a maximum): "one would have supposed that in doing this the investigator would eventually approach a position which was identical with the criterion-axis, and that the maximized correlation . . . would be simply the ordinary multiple correlation between the criterion and the tests." This quotation shows that the reviewer has failed to understand the theory of criterion analysis. There can be no multiple correlation between the criterion and the tests, because the criterion is not included in the factor analysis, and cannot be so included, by virtue of the fact that the analysis is carried out *separately* on the normal and the abnormal groups, *not* on the *combined* groups. This is the most important point in criterion analysis. If the reviewer has failed to appreciate the consequences that follow, it may be legitimate to query whether his understanding is adequate to the task he has undertaken.

9. *Misrepresentation through carelessness.* He follows the criticism quoted above by saying: "It is true that he (Eysenck) goes on to say that one may also increase the correlation by improving the criterion. . . . How this is to be done we are never told." The reader is told at length on pages 76 and 77. W. L. G. would be justified in arguing that he did not agree with the arguments—or did not understand them. He cannot plead that they are not to be found.

10. *Misdirection through non sequitur.* The reviewer quotes very briefly my account of the hypothetico-deductive method, as involving the postulation of an hypothesis, the deduction of certain consequences, and the verification of these consequences along experimental lines. He then declares it to be invalid by quoting Judge Jeffreys' saying: "If Mistress Gaunt had been guilty, you may rest assured she would have been burned. She was burned. *Ergo*, she was guilty." In some obscure way, he seems to feel that this story is relevant, and disproves my own argument. This seems to me a complete *non sequitur*. I leave it to the reader to decide whether he regards this as a telling and relevant argument.

11. *Misrepresentation through misstatements of facts.* In connexion with my discussion of scientific methodology he says: "To prove his (Eysenck's) hypotheses, it is not sufficient to report experiments that are consistent with his corollaries: *he should also have attempted to disprove the alternatives.* When he does notice alternatives, they are usually dealt with by the short and easy method of 'abusing the defendant's attorney.'" When we look at the facts, we find that the main theory tested, namely, that of the qualitative or quantitative nature of abnormal mental states, as compared with normals, is examined by means of a crucial experiment, i.e., an attempt is made to deduce consequences from either theory which would be incompatible with consequences deducible from the other. The critic may argue that the attempt was not successful; he can hardly remain consistent with the facts of the case and maintain that I never even made the attempt to disprove the alternative hypotheses. Similarly, I do not try to invalidate current psychiatric diagnostic systems by abuse. I deduce certain consequences (such as inter-psychiatrist reliability) which must follow from a satisfactory system, show that these consequences in actual fact are not to be found, and conclude that the original hypothesis is not tenable in its present form. The arguments may be faulty; but it is contrary to fact to say that there is 'nothing but abuse.'

12. *Use of alternative plea.* The alternative plea is very useful in law, where a man may plead that he did not kill the victim, but that, if he did, he did it in self-defence. Scientific criticism tends to frown on the use of this device, which is exemplified in its purest form by my reviewer. On p. 210 he assures the reader that my treatment of abnormal mental states as quantitatively different from normal people is opposed to Burt's assumptions and methods; on p. 211 he assures us that the two systems and the concepts involved are similar, and that I have in effect taken mine from Burt. But, while either one or the other of these arguments may be true, they cannot both be true. There are other instances of this peculiar type of argument in the review; it would hardly be worth while to exhumate them.

13. *Misrepresentation through wrong emphasis.* Throughout his review, the writer attempts to relate my results to those of Burt, castigates me for not agreeing with Burt, attempts to establish Burt's priority with respect to such concepts as *emotionality*, and quite generally writes as if no other psychologist but one of the editors of the *British Journal of Statistical Psychology* had ever written on the subject of personality. This emphasis is erroneous in two ways. In the first place, my book is an account of experiments, not an historical treatise: occasionally previous writers are mentioned, but there is no systematic attempt to cover the ground. This was originally contemplated; but the material was too extensive, and will soon appear as a separate publication under the title *The Structure of Human Personality* in the Methuen series. Criticisms of an historical character are appropriate to this later volume: they are irrelevant to the present one. In the second place, the criticisms made are poorly taken. If there are similarities between Burt's system and mine, which I would be the last to deny, they presumably are due largely to the fact that the early workers (Jordan and Gross on the theoretical side, Heymans and Wiersma on the experimental side) laid down the

foundation on which we both have built. I had already commented on the relation between my system and Burt's in *Dimensions of Personality*; there seemed little point in discussing it again. The reviewer seems to regard this lack of repetition as proof that I have changed my mind—a curious hypothesis also put forward by Sir C. Burt. In fact, the very term “emotionality,” with almost exactly the connotation which it carries in Burt's system, was used by Heymans and Wiersma in 1909; and the existence of the trait so denoted verified by an empirical procedure closely analogous to factor analysis. If, therefore, I should have entered into the historical side, the correct emphasis would not have been on Burt, but on these other, earlier workers who anticipated his system (and mine) in considerable detail. But it is not with respect to the system itself that I would claim to have made any great advance: my aim has been rather that of improving the underlying methodology and making more precise the measurement of the dimensions posited.

Here, then, are a baker's dozen of serious misrepresentations of fact, of misquotations, of errors, of misunderstandings, and of complete inventions having no counterpart in the book under review. The list could have been extended, but it is perhaps long enough to cast doubt on the reviewer's knowledge of the field, his ability to understand the complex arguments involved, and his impartial fairmindedness in assessing the validity of proofs and demonstrations.

THE SCIENTIFIC STUDY OF PERSONALITY: A REPLY

Methodology. In answering Dr. Eysenck's thirteen criticisms I shall assume that readers will be more interested in the appraisal of his own methods and conclusions than in personal questions about my competence as a psychiatrist or psychologist, which, I modestly admit, is sadly defective. If I have in any respect misrepresented or misquoted his views, I offer at once my sincerest apologies. Nevertheless, a reference to what I actually wrote will, I think, show that he has misquoted me far more often than I have misinterpreted him.

He has told us that the chief aim of his book was to “improve the underlying methodology” in the field of personality research. It was with this that my review was chiefly concerned; and it is with this that my reply will chiefly deal. What he describes as ‘incomplete quotation,’ ‘misrepresentation through suppression,’ ‘non-sequiturs,’ and the like, arose mainly from my efforts to be brief. Hence the best way to meet his complaints will be to amplify the points over which misunderstanding seems to have arisen.

The Verification of the Proposed Hypothesis. I began, as he says, by “quoting very briefly” the three steps in his ‘account of the hypothetico-deductive method’—‘postulation,’ ‘deduction,’ ‘verification.’ So far, I gather, there was no ‘distortion.’ But “then,” according to him, I “declare it to be invalid” by just narrating a fallacious syllogism attributed to Judge Jeffreys. “In some obscure way,” says Dr. Eysenck, “the reviewer seems to feel that this story disproves my (Eysenck's) argument.” If, however, Dr. Eysenck will read the paragraphs and the footnote that follow “this story,” he will see that it was inserted merely to *illustrate* the formal nature of the argument, not to “disprove” it. I went on to indicate quite explicitly why such a mode of inference was defective, namely, that it “commits the fallacy of affirming the consequent.” In deductive logic the only way to prove a hypothetical premiss is to disprove its contradictory (Stebbing, *loc. cit.*, p. 105). Instead of elaborating the point at greater length, I thought it sufficient to refer to several logicians and statisticians who had already discussed it, and whose objections Dr. Eysenck had ignored in his book, and still ignores in his reply.

In his chapter on the subject, instead of dealing with the criticisms so often raised, he preferred to appeal to a well-known authority, and for this purpose introduced the long quotation from Miss Stebbing's text-book (*A Modern Introduction to Logic*, 1930). But this mode of defence suffers from two grave weaknesses. First, Miss Stebbing's chapters can hardly be regarded as the last word on statistical methodology. Secondly, her own discussions do *not* support his view; and there can be little question that Dr. Eysenck has seriously misrepresented, or at least misunderstood, her real opinion.

If the reader will refer to the original, he will be amazed to find that Dr. Eysenck, who now lays such emphasis on exact quotation, has run together in one continuous extract sentences and passages scattered over many pages of her book, rearranged their order, reworded several of the statements, and even inserted a sentence of his own.¹ “The method of science” (he makes her say, though she says no such thing) . . . “has four essential stages”; and then he gives her analysis, not of ‘the method of science,’ but of ‘common sense thinking.’ She herself points out that an induction proceeding by these four stages is invalid, since “a formal fallacy is involved in the inference that a given hypothesis is true because the consequences deduced from it have been verified”; and explains in some detail how the “method of science” improves on this inconclusive procedure in several “important respects.”

¹ If the reader will compare pages 227 to 242 of Miss Stebbing's book with pp. 16–18 of Dr. Eysenck's he will find excellent instances of ‘misquotation,’ ‘incomplete quotation,’ ‘misrepresentation through suppression,’ ‘wrong emphasis,’ ‘invention,’ etc., etc., all in a single passage. I will not go so far as Dr. Eysenck, and suggest that it offers “an almost perfect example of the art of distortion”; but it certainly presents a very misleading picture of Professor Stebbing's own analysis of inductive methods as applied to the physical and the biological sciences.

The Disproof of Alternative Hypotheses. The obvious way to escape any such formal fallacy would be to attempt a direct disproof of all the conceivable alternatives to whatever hypothesis the writer wishes to maintain. In empirical science, however, the most we can seek will be, not deductive proofs leading to conclusions that are *certain*, but arguments based on probability, leading at best to an inference that the hypothesis favoured by the investigator is more *probable* than the remaining alternatives.¹ I thought it unnecessary to amplify these points in any detail because they had already been adequately dealt with in the references I quoted.

But although in section 10 of his reply Dr. Eysenck seems to reject my general contention, in section 11 he apparently admits it. He now claims that, "for the main theory tested," he did in point of fact "disprove the alternative hypotheses." But he interprets this requirement in a way very different from most other methodologists. Hence I still think my two chief criticisms were justified. First, the need for this all-important step is nowhere mentioned in his account of the 'hypothetico-deductive methodology.' Secondly, in what are described as the central enquiries of the book, there is no *direct* attempt to disprove the alternatives. We are, it is true, told that, where Kretschmer holds a different view, his methodology is 'faulty' and his attempts at proof 'disingenuous.'² For the rest the disproof of the chief alternative is treated as an incidental consequence of the proof of the main hypothesis, whereas the correct mode of reasoning would invert the order. However, in view of the way in which he seems to have missed my principal point, let us examine a little more closely his study of psychosis, since this is the part of his book which he complains that I ignored, and the part which has chiefly aroused criticism from medical writers. Here he claims, by means of the hypothetico-deductive procedure, to have established two far-reaching conclusions.

Hypothesis I. The first is the hypothesis that "mental states diagnosed as psychotic by psychiatrists" are merely instances of extreme variation in a single *quantitative* continuum—the 'dimension of psychoticism,' which ranges from perfect normality to definite psychosis. The alternative would be the view that certain psychoses at least differ *qualitatively* from each other and from the normal state, much, for instance, as Korsakow's psychosis is ordinarily supposed to differ from general paralysis, and both from normality. In accordance with the steps described in his opening chapter, Dr. Eysenck first deduces what he takes to be the consequences of the 'quantitative hypothesis' (that of continuity), and then attempts to verify these. But, so far as I can discover, he deduces no consequences from the *alternative* hypothesis, nor does he undertake any independent investigation to disprove them.

To begin with, he puts all his cases of psychosis³ into a single group. No attempt is made to show that his '100 psychotic subjects' were typical psychotic patients; indeed, from their excellent performances at the tests most psychiatrists would be tempted to suppose that they were highly selected cases. He then argues that, if the psychotic group is continuous with the normal, tests which differentiate significantly between the two should yield *positive inter-correlations for each group separately* (his italics). He adds that these inter-correlations should be "proportional to the power of each test to differentiate between the normal and psychotic groups," the 'differentiating power' being measured by a biserial correlation.⁴ He offers no formal demonstration that this corollary follows logically from his hypothesis: he merely "points it out." The coefficients obtained from his tests are tabulated in Tables XL and XLI; and he thereupon concludes (p. 219) that "we find our deduction verified, and it would appear that psychotic states do in fact

¹ It is in regard to the relation between induction and probability that Miss Stebbing's section on 'The Problem of Scientific Method' seems somewhat out of date. The reader may usefully compare both her views and Dr. Eysenck's with those put forward in Russell, B., *Human Knowledge*, 1948, esp. Pt. VI, 'The Postulates of Scientific Inference.'

² In his *Note* Dr. Eysenck ends his section on this point by declaring: "it is contrary to fact to say that there is 'nothing but abuse.'" But I said no such thing. What I did say was that, in discussing alternative theories, he seemed a little too apt to "abuse the plaintiff's attorney." I used this familiar cliché to mean that, instead of adducing facts to disprove the views he dislikes, he relied rather too freely on an attempt to disparage the scientific competence of those who advocate some alternative conclusion.

³ In attempting to assess the real implications of Dr. Eysenck's statistical studies, the practical psychiatrist is rather at a loss owing to the inadequate description of the patients tested. We are simply told that his "psychotic subjects" (as he calls them) were "divided into two equal groups consisting of 50 manic-depressives and 50 schizophrenics." No details are vouchsafed to show whether the individuals so labelled were typical of the highly heterogeneous conditions to be found under these somewhat vague diagnoses. I should have thought that by now even the statistical psychologist was convinced of the need for combining case-studies with mere quantitative analysis.

⁴ His use of biserial correlation to measure the 'discriminative power' of his tests presupposes that the frequency distributions for the combined group (e.g., the 100 normals and the 100 psychotics) would be, not only continuous, but also normal (as shown in his schematic diagrams). Now, if we reduce the two constituent groups to appropriate proportions, we can test the assumption of normality by applying Pearson's equations for the effect of selection on the correlations and standard deviations. So far as can be judged from the data Dr. Eysenck gives, it seems clear that the hypothesis of a normal distribution along the single continuum is far from being fulfilled.

form a continuum with normal." A few paragraphs later he adds: "we have shown" that psychotics are not, "qualitatively different from normal people," i.e., that the alternative hypothesis is disproved. But the only evidence for this further inference is a parenthetical assertion (unproved) that the consequences of his own hypothesis "would not follow on any tenable counter-hypothesis." Apparently he assumes that with any 'counter-hypothesis,' e.g., the hypothesis that differences were qualitative, the average inter-correlation would be zero. Yet this is never proved; and is it really so? Suppose we were investigating the analogous alternatives in a study of sex-differences: would Dr. Eysenck expect to find that if sex differences were qualitative (due, say, to an additional X-chromosome), then the inter-correlations between measurements for height, weight, shoulder breadth, etc., would yield an average correlation of zero in males and females taken separately?

Or consider the parallel question that arises in the field of cognitive psychology: do patients in mental hospitals who have been diagnosed as imbeciles or idiots differ from normal persons solely because of some "quantitative variation in a single continuum"? Take any two tests that will 'differentiate significantly' between 100 normals and 100 defectives, say, any pair from Abelson's battery; and calculate their correlations within each group taken separately. The coefficients will undoubtedly be, not zero, but positive and significant. On taking all the tests in the battery, we should find that the average of the correlations for each (and therefore its factor-saturation) would vary in close proportion to its 'power to discriminate' between the normals and the defectives. Dr. Eysenck would thereupon conclude that the results had shown that "neither [idiots] nor [imbeciles] are qualitatively different from normal people." Nevertheless, among institutional cases of imbecility, as Professor Penrose assures us, while a small proportion display merely an extreme quantitative variation in general intelligence, as many as 75 to 80 per cent. suffer from some "chronic neuropathological condition," due to some specific gene, or to some disease, deformity, or qualitative defect (often quite specific) without which they would have had "average intelligence."

The fact is, as an algebraic proof would quickly show, that Dr. Eysenck's assumed 'deductions' do not follow from his hypotheses as he states them. In any case, his decision to rely almost exclusively on graded tests scarcely affords a genuine chance of proving or disproving the presence of a qualitative difference. In explaining his choice of test procedures, he asserts that individuals "cannot be meaningfully said to differ unless they are compared along some quantitative variable." But, as I went on to argue, this is tantamount to assuming the very continuity that he later claims to demonstrate. In reply he now protests (section 6) that he did not make any assumption about what he would find. Naturally I did not mean that he deliberately or consciously did so. Yet at times he certainly speaks as if the main object of his investigation was merely to confirm, by a quantitative factorial technique, an assumption which, on more general grounds, he took to be almost a foregone conclusion from the start. Thus, in his preliminary remarks on 'the dimensional approach,' he tells us that the whole notion that persons can be sorted into qualitative categories—a view "often maintained by some psychiatrists"—really rests on an obsolete type of thinking. "Such thinking," he says, "has always characterized early stages of scientific progress: usually it is shown later that what was considered qualitatively different is reducible to quantitative variation along a continuum."¹ No doubt, as he contends, unless some such postulate is accepted, factor analysis as he understands it could hardly be applied; and (so he maintains) "at the present stage of our knowledge there does not exist any other method which could aid us in our quest." But that is a somewhat sweeping and surprising pronouncement; moreover, it is a pronouncement which (as I said in my review) most psychiatrists would assuredly deny. There are many other methods available, observational, experimental, etiological, and clinical, as well as factorial and statistical. And even among factorial and statistical procedures there are several which would appear to be just as appropriate for qualitative characteristics as Dr. Eysenck's are for quantitative variables.

Hypothesis II. I now turn to the second of the two main hypotheses about the nature of psychoses which Dr. Eysenck has discussed. As he points out, the views of Kretschmer and other psychiatrists imply the presence of at least one further 'dimension' or factor, namely, "a continuum running from the extreme schizophrenic to the extreme cyclothymic" (or manic-depressive). Dr. Eysenck's twenty tests are designed to investigate this second hypothesis as well as the first. Their inter-correlations do, as a matter of fact, yield two factors. Of these the first is identified with the factor of psychoticism; and therefore, it is argued, the second should be identical with the dimension contrasting 'schizophrenics' with 'depressives,' if such a dimension exists.

¹ Here it is surely Dr. Eysenck who is basing his own generalization on views that are "thirty years out of date." What he says is not true of other sciences. "The idea of a continuum seems so simple to us," says Schrödinger, "owing perhaps to conditioning in early childhood. . . . If, however, we envisage the development of physics during the last half century, we get the impression that the discontinuous aspect of nature has been forced upon the scientist *much against his will* (Schrödinger's italics): he seemed to feel so happy with a continuum." (*Science and Humanism: Physics in our Time*, 1951, pp. 30, 54f.) The same is true in the biological sciences, notably in the genetic study of human abnormalities. "The view that ascribed all differences to continuous variation has got to be abandoned: what is fundamental are qualitative differences, not quantitative" (cf. Myerson, A., *The Inheritance of Mental Diseases*, 1923).

Now the saturations for the second factor reveal a zero correlation with the biserial correlations which measure the power of the tests to discriminate between 'schizophrenics' and 'depressives.' Therefore Kretschmer's notion of a second psychotic factor is disproved: "it would appear that schizothymia-cyclothymia does not exist as a separate dimension of personality." Moreover, since the scores of the schizophrenics are in general somewhat better than those of the manic-depressives, we may (it is suggested) reasonably conjecture that schizophrenia is, as it were, a mild form of manic-depressive insanity, intermediate between the latter and perfect normality.

But surely there are other ways of accounting for the zero correlation and the scores. First of all, just because Eysenck's second factor and Kretschmer's second factor are both orthogonal to psychoticism, it does not follow that the two must be one and the same; several dimensions or factors might exist,¹ all orthogonal to psychoticism and to each other. Secondly, the saturations for Eysenck's second factor are of very doubtful value. They are reached after a double reversal of signs; and, whenever the initial coefficients are small, such reversals are extremely arbitrary: with a different mode of reversal, quite a different correlation with the biserial coefficients might have been obtained. In any case, the figures correlated have for the most part a decidedly low degree of significance; of the biserial coefficients, at least 16 are definitely non-significant; of the factor-saturations none of the negative figures in the normal group is significant, and two at most in the psychotic group. Thirdly, it is quite possible that such tests as 'reading prose,' 'abstractions,' 'mirror drawing,' and the like—even if their reliability remained reasonably high when performed by psychotic patients—were not suited to elicit the special characteristics of the schizothymic and cyclothymic temperaments. In view of the similarities between schizothymic tendencies and introversion, and between cyclothymic tendencies and extraversion, I should have thought that the methods adopted with some success by previous investigators in their endeavours to verify the factor of introversion and extraversion might have furnished more convincing evidence. The mere fact that "the requisite deduction could not be verified" hardly justifies the far-reaching inference that is drawn. Finally, the test results themselves seem highly inconclusive. Dr. Eysenck's first and most important factor ('psychoticism') only accounts for about 10 per cent. of the test-variance. His second factor is smaller still. Indeed, he himself does not venture to say what it really represents.

I do not doubt that most of the mental characteristics which Dr. Eysenck has tested, and many of those which, in an exaggerated degree, form the more superficial symptoms of the neuroses and milder psychoses, might be expressed in terms of quantitative factors of the usual type. But that is by no means the same as "showing that neither neurotics nor psychotics are qualitatively different from normal people." The temperature, rate of pulse, rate of breathing, and other measurable symptoms of patients suffering from measles or typhoid fever could be expressed in terms of one or two quantitative factors. But that does not demonstrate that there are no qualitative differences between measles and typhoid, or between either of these and normal health.

Reliability of Tests. In my review, I ventured to suggest that possibly the low coefficients which Dr. Eysenck has obtained—biserials, inter-correlations, and saturations—might be attributed to the fact that the types of test chosen were not likely to yield very reliable measurements, particularly with neurotic and psychotic patients, and I thought it unfortunate that Dr. Eysenck's tables "seldom give reliabilities." In his note he replies that this is "disparagement by suppression." Nevertheless, in his thirty tables I can still discover only two which mention 'reliabilities,' and these relate to researches carried out by other investigators. He now tells us that four of the tests he himself has used (body-sway, persistence, and the like) "are reported in the literature to have satisfactory reliabilities"; but he gives no references. The references that I myself suggested (which give rather low reliabilities for tests of the general type in question) he dismisses as "onesided." In any case, he says, the reliabilities for his own tests (or rather their lower limits) could be inferred from the validity coefficients. But with the twenty tests used to discriminate psychotics from normals the highest validity coefficient was 0.402, and with those used to discriminate neurotics from normals none reached 0.600: the averages are about half the size. Such values can scarcely be claimed as proving that the reliabilities are "satisfactory."

Dr. Eysenck roundly declares that I have "suppressed the fact that these tests have been shown to be valid, and must therefore be reliable." I did not 'suppress' the fact; I questioned whether

¹ In accordance with the principles of 'criterio-analysis,' Dr. Eysenck also examines the correlations obtained for certain factors reached by rotation. Once again he finds them non-significant; and therefore considers his conclusion to be confirmed. But in any case mere rotation could not possibly discover any new dimension.

In section 7 of his reply, he complains that I wrongly supposed that his method of rotation was graphical, whereas it was "precise and algebraic." If that was a mistake, I apologize. However, in the initial exposition of the proposed procedure (p. 73), he says nothing about an 'algebraic' rotation, and he does speak of "a graphic demonstration of the results." He dismisses my remarks on 'criterion-' or 'criterio-analysis' (as he variously terms it) by saying that my 'understanding (of it) is inadequate.' But my main criticisms were really those urged by his former collaborator Dr. Lubin (*Psychol. Rev.*, LVII, pp. 54f.), who must surely have understood the procedure. To these Dr. Eysenck still attempts no reply.

it was a fact. And at this point a critical reader might be tempted to ask whether it is not Dr. Eysenck who has 'suppressed' some of the relevant facts. In his book (p. 125) he states that the "testing was carried out by Drs. Clarke and Gravely" (who at that time were co-operating with him at the hospital), and that "the account given by Clarke has been followed . . . often verbatim." But Dr. Clarke tells us that "none of the tests" (which included "body sway" and others "selected in consultation with Dr. Eysenck as the most promising") "has a sufficiently high validity coefficient to claim any practical value for screening or diagnosis" (*Brit. J. Educ. Psychol.*, XX, p. 202). Dr. Gravely reports very similar results: (see the theses quoted by Dr. Eysenck). These findings scarcely support his present line of defence.

Priority. Dr. Eysenck's last and longest section is devoted to the question of priority. He complains that "throughout [my] review [I] attempt to relate [his] results to those of Burt, castigate him for not agreeing with Burt, attempt to establish Burt's priority with respect to such concepts as 'emotionality'" (a concept which, he says, was really due to Heymans and Wiersma), and generally that I "write as if no other psychologist had ever written on the subject of personality." The last sentence is surely an instance of what Dr. Eysenck has labelled "misrepresentation through misstatement of fact." If he will be good enough to refer once again to my review, he will see that I refer to more than a dozen other psychologists by name, including "eminent writers on personality, like Allport . . .," and several others whom Dr. Eysenck himself had ignored. I made no attempt to establish anyone's priority in regard to the concept of emotionality. The sole priority with which I was concerned was priority relative to Dr. Eysenck, and that in respect of his general scheme of factors, not in respect of one particular concept. Nor did I "castigate" him for *not* agreeing with Burt; I simply regretted that, on the numerous points about which he *did* agree, not with Burt merely but with other investigators, the only reference he generally gave for the "demonstration" of the several factors was, as a rule, some particular paper of his own. He replies that his work was not meant to be "an historical treatise." But I was not asking for a history, but simply for a less misleading mode of reference. His 'bibliography' runs to more than 300 numbers; yet this, like the text, omits most of the previous British investigations which have dealt with the factors he describes and reached much the same conclusions. I had presumed that the omissions might be due to the fact that he was unfamiliar with the earlier English literature. He rejoins that apparently I did not "consider the possibility that other causes than ignorance might be responsible." I did; but thought it more charitable to dismiss them.

He professes surprise that I should relate his scheme to that of Burt. There are two obvious reasons. First, apart from the single factor of psychoticism and a slight change in nomenclature, the two schemes seem almost identical. Secondly, Dr. Eysenck received his own psychological training under Burt, and carried out his earlier factorial researches in his laboratory. Indeed, in his previous book he had proposed "a theory of personality organization," which, he said, was "based on Burt's views of the nature of the statistical factors and on the hierarchical theories of McDougall."

However, he now explains that, "if there are similarities between Burt's system and his," they must be due chiefly to the fact that both he and Burt built on the foundations laid down by early workers like Jordan, Gross, Heymans and Wiersma. As early as 1909, he tells us, Heymans and Wiersma used "an empirical procedure" which (in his view) is "closely analogous to factor analysis" in order to demonstrate the existence of hypothetical traits like emotionality. This explanation of the resemblances, I confess, is somewhat unexpected. Nowhere in his book is any of these four names mentioned. Nor do I know anything to support the view that Burt derived either his methods or his hypotheses from the work of these four writers rather than from that of Wundt, McDougall, and Pearson. In point of fact, as a glance at the dates will show, his first investigation with factorial methods was started under McDougall some two or three years before the date given by Dr. Eysenck for the publication by Heymans and Wiersma. Dr. Eysenck's first investigation on the subject (based, as he expressly states, on Burt's factorial technique) was not published until 30 years later.

All psychiatrists would readily admit that, at any rate until quite recently, German contributions to the study of mental disorders had been well in advance of British. But I venture to think that Dr. Eysenck's enthusiasm goes a little too far when he claims that the origins of the factorial study of personality are also to be found in the *Zeitschrift für Psychologie*. Nevertheless, it is well to be reminded of the work of early continental writers, which British writers have perhaps too often overlooked.

Misquotation. The rest of the thirteen charges in Dr. Eysenck's long indictment will perhaps seem to the reader too trivial to spend time over. However, since they have been pressed so strongly, it is only right that my answers should be briefly recorded. The more important relate to alleged misquotations. But in most cases the misquotation appears to be Dr. Eysenck's rather than mine. Thus, in section 2 he refers to 'Burt's 1915 Report' describing the general and specific emotional factors discovered by analysing correlation tables; and in commenting on his account I had stated that the coefficients were "based, not on impressionistic ratings obtained from teachers, but on items of behaviour systematically recorded." But here, he says, I am guilty of misquotation, since Burt "does not tell us on what his correlations were based." If so, I apologize to Professor

Burt; but the '1915 Report'¹ explicitly states that the investigation "departed from the method hitherto used in studies of character qualities" and that the assessments correlated were based on items of behaviour "systematically observed and recorded for all members of the groups." It is therefore Dr. Eysenck whose account conveys 'a completely erroneous impression.'

In section 3 he gives yet another example where my method of quotation is said to distort what was meant. On page 115 of his book Dr. Eysenck had asserted that, in a chapter contributed to the American Symposium, "Burt has . . . contrasted the factor of 'emotionality' with the factor of 'neuroticism.'" I ventured to correct this, and observed that Burt's real contention was that the two are "virtually identical." Dr. Eysenck now tells us that my quotation was inaccurate and broke off too soon. But, as a matter of fact, it is Dr. Eysenck who breaks off the quotation too soon. Burt's essential point is that the difference lies, not in the factors themselves, but only in their names. He cites an enquiry carried out among both neurotics and normals which showed that the so-called general factor of 'neuroticism' was not confined to neurotics only, but was also found among those who were perfectly healthy; and from this he concludes that the factor is "little else than the factor of general emotionality."² Dr. Eysenck, in quoting this paragraph, omits this concluding sentence.

Although throughout his *Note* Dr. Eysenck directs his criticisms mainly to Burt's supposed opinions, it should be observed that many other factorists³ have come to much the same conclusions as those expressed in my review. It is thus Dr. Eysenck who writes as though only one psychologist of repute had put forward opinions differing from his own. Indeed, he goes out of his way to ascribe every antagonistic view to the same single source, often (to borrow his phrase) by "subtly changing" what I actually wrote. In section 12 of his *Note*, for example, he contrasts my discussion of his 'hypotheses' with my discussion of his 'conclusions,' and alleges that at the beginning of the former I "assured the reader that [Eysenck's] treatment of abnormal mental states as quantitatively different from normal people is opposed to Burt's assumptions." But I said nothing at all about Burt's assumptions in regard to adult psychotics, which, for all I know, may be much the same as Dr. Eysenck's. What I said was, "many psychiatrists" would be opposed to Eysenck's assumptions, and would probably reply that such differences "were to be expressed in terms, not of a single quantitative variable, but of qualitative attributes." I supposed that anyone familiar with the literature of this controversy would know that I had in mind such views as those put forward most convincingly perhaps by Eliot Slater in various papers (e.g., 'The Neurotic Constitution,' *J. Neurol. Neurosurg. Psychiat.*, VI, 1943, pp. 1-16, 'A Heuristic Theory of Neurosis,' *ibid.*, VII, 1944, pp. 49-55) and tacitly or explicitly adopted by all but a few psychiatric text-books.

Let me add that, although I have ventured to criticize the accuracy of his arguments, this does not mean that I myself necessarily agree with the views that he has rejected. Moreover, I hope that his numerous complaints about the way he supposes I have 'disparaged' his work will not blind readers to the fact that my review went on to express a sincere admiration of "its undoubted merits"; and I more than once paid a tribute to the "ingenuity and industry it displays."

What I most of all regret in his reply is that, by concentrating on these rather trifling points about the incompleteness of the arguments or quotations in a review that was already too long, he has diverted attention from the more important issues I endeavoured to stress. I had hoped that the questions I raised might serve to clarify the conflict between Dr. Eysenck and the psychiatrists (to which Professor Lewis refers in his Foreword) and perhaps elicit a fuller and clearer explanation

¹ I take it that this refers to his 1915 L.C.C. *Report*. But the method used is also indicated in the brief abstract of his British Association paper. If Dr. Eysenck still has doubts on the matter, he will find the experiment fully described in Burt's later article in *Character and Personality*, and the method itself is again explained in *Brit. J. Med. Psychol.*, XVII, p. 164. Let me add that, in trying to meet Dr. Eysenck's criticisms and verify his quotations, I have found myself constantly handicapped by the careless way he gives his references. In his bibliography, for example, the only contribution given for Burt under the year 1915 is a paper read to the British Association: but, if one turns up the volume and page specified by Dr. Eysenck, one finds that there is nothing whatever by Burt on that page at all. Passing to the next reference to discover which of Burt's 1937 papers is being criticized in section I, the reader finds that the only article cited under that year is one which was not published in that year. And there are many similar slips which cannot all be due to the printer.

² *Feelings and Emotions* (ed. M. L. Reymert), 1950, p. 548. Elsewhere Burt had contrasted the so-called general factor of neuroticism with the various specialized factors for certain types of neurotic disorder. The former he believes to be wrongly named, and to have diverted attention from the latter, which, he suggests, are from a psychiatric standpoint far more distinctive (see this *Journal*, I, pp. 193f.).

³ See the numerous references cited by Raymond Cattell in his *Description and Measurement of Personality* (Index, s.v., 'General Emotionality Factor' and 'Neuroticism.'). Cattell himself writes: "Probably the worst example in applied psychology of confusion through inadequate factoring occurs with respect to the alleged 'neuroticism' factor; later studies show half-a-dozen such factors, and the first has no more claim than the others to be 'the neuroticism factor.'" He himself identifies it with his 'factor C.' This, he says, 'ties up with general emotionality,' but is as important in the causation of delinquency as in the causation of neurotic disturbances: (cf. also *id.*, *Personality*, p. 502).

of the statements that had aroused some criticism. Unfortunately in his *Note* as in his book Dr. Eysenck merely waves away the views of those who differ from him as being 'one-sided,' 'valueless,' or 'based on lack of understanding or knowledge of the field'; and we still have no real reply to the questions themselves.

In criticizing his 'attempt to improve the underlying methodology,' my real objection was not so much that his factorial technique might be faulty in this or that detail, but rather that a *purely* factorial approach is of itself very liable to oversimplify the facts. What I was pleading for was a more eclectic 'methodology,' like that, for example, which has proved so successful in "unravelling delinquency"—a set of co-operative researches which should combine the method of case-study with that of statistical analysis, and the method of clinical observation with that of the laboratory test. And since Dr. Eysenck seems to imagine that I am just an isolated psychiatrist, unfamiliar with the requirements of a strict scientific or statistical procedure, and that my "knowledge of the field" is confined to just one brand of factorial psychology, let me add that, since I wrote, others have brought forward much the same criticisms.¹ With a far better knowledge both of the opportunities and of the difficulties, they have also urged what I ventured to hint at towards the close of my review: namely, that, placed as he is on the staff of one of the best-equipped mental hospitals in the country, it seems disappointing that he has not been able to make fuller use of the unique facilities thus afforded for linking up his tests and quantitative data with the first-hand clinical observations made by the skilled psychiatrists who form his colleagues and who presumably had charge of the psychotic patients he has tested. What might be achieved along these lines has already been foreshadowed in the papers contributed by the little band of psychologists and psychiatrists who worked together before the war at the same mental hospital under the guidance of Professor Spearman. And I should like to echo the remarks of Dr. Stephenson, one of the most active members of that group: "by themselves factorial studies can serve only the broadest aims; conjoined with case-records, each can help to interpret the other, and together reveal common characteristics and common trends."

W. L. G.

LABORATORY NOTES

To the Editor, British Journal of Statistical Psychology

SIR,—In previous publications in your *Journal* and elsewhere I have seen references to 'Laboratory Notes' on (1) the Psychology of Humour and (2) the Content Method of Scoring the Rorschach Tests. If copies are still available, we should be grateful to know how they can be procured. If not, would it be possible to have some indication of the contents?—Yours, etc., E. A. TELFORD.

Reply.—A few copies of the Notes are still available, and can be obtained by writing to me at the Department of Psychology, University College, Gower Street, London, W.C.1. The note on 'Humour' includes a fairly complete summary of the theories of humour, and a report of illustrative investigations carried out in the laboratory (down to 1936, when the note was prepared) dealing with the more conspicuous points of disagreement. Tables of correlations are given both 'between items' and 'between persons,' together with the results of the factorial analyses (based mainly on work by Pelling, Dewar, and Woods). The note on the Rorschach Test tentatively puts forward a detailed scheme for deriving trait assessments from a study of the contents of the replies rather than of their form (see this *Journal*, III, pp. 24f.).—CHARLOTTE BANKS.

PUBLICATION OF TABLES

SIR,—I venture to suggest that authors who wish to contribute statistical papers to any of the journals published by the British Psychological Society should be required to supply copies of the essential tables on which their inferences are based. If possible, the tables should be printed with the article. If (for reasons of expense, space, or lack of general interest) this is impracticable, then the copies might be deposited for reference in the Society's office.

To illustrate the difficulties arising when tables are not available, I may refer to a recent paper on retardation among school pupils, published in another journal. The author gives no tables of correlations or saturations; but describes in some detail the factors obtained apparently by some method of rotation. We happen to be carrying out a research in the same field at the Queen's University, Belfast; and in this case I was very courteously allowed to consult the unpublished thesis. On examining the tables, it seemed clear that another interpretation might easily be placed on the loadings obtained. Other readers might desire to compare these factors with those published by Barakat, Sutherland, and other authors in this *Journal*. And in many cases it might be difficult, if not impossible, to obtain access to such theses. I therefore wish to put forward this proposal as a general policy for all editors who publish articles based on statistical data.—Yours, etc., JACK WRIGLEY.

¹ Dr. O'Connor perhaps has put the matter most emphatically. "Dr. Eysenck's false god," he writes, "is the statistical table . . . so pleasing to that kind of scientific mind which prefers to remain aloof from the human element" (*Bull. Brit. Psychol. Soc.*, III, 1952, p. 115). Dr. Maddox has criticized the tests employed even more severely than I had done; and several later reviews, I find, raise much the same objections as I had.

BOOK REVIEWS

Contributions to Mathematical Statistics. By R. A. Fisher. New York : John Wiley and Sons, Inc. London : Chapman and Hall. 1950. Pp. 680. £3.

This large volume contains the first comprehensive collection of Sir Ronald Fisher's writings on mathematical statistics, and forms a notable addition to Wiley's series of statistical publications. The selection has been made by the author himself. Forty-two contributions have been taken from articles and lectures published before 1943 ; a further note refers to three others, not actually included, of which one appeared as early as 1915. The papers, some of which were printed in octavo journals, others in quarto, have been reproduced on their original scale by a photographic process. Professor Fisher has made a number of deletions and emendations, and provided brief introductory notes, linking the various contributions with each other and setting them in the context " of the common knowledge, or common misapprehensions, current at the time they were written." An index to the whole has been prepared by Professor S. S. Wilks.

The modern theory of mathematical statistics has been almost entirely created during the last seventy years. The *Contributions* thus cover nearly half this period. Although in many respects they form a fairly self-contained body of knowledge, it is scarcely possible to understand the achievement they represent, except in relation to statistical developments during the earlier half of the same period. This witnessed the appearance of the long series of publications by Professor Karl Pearson and his ' biometric school ' ; and it might be said that the place held in the earlier phase by Pearson's ' Mathematical Contributions to the Theory of Evolution ' and kindred papers is as central as that held by Fisher's *Contributions* in the later. Under Pearson's leadership, detailed investigations were made of frequency functions ; these led to the well-known Pearsonian classification of frequency curves and the use of moments for fitting them to empirical distributions ; most of the multivariate descriptive procedures appropriate to biometry, including nearly all the familiar coefficients of correlation and regression, were systematically developed ; and the problems of sampling, and of significance-tests as applied to large samples, were studied both experimentally and mathematically. The chi-squared distribution—the first of the more modern frequency distributions—was rediscovered, and has since become closely associated with Pearson's name. On the other hand, the nature of statistical inference was very imperfectly understood. The method of inverse probability, with Bayes' postulate, was accepted as the foundation ; routine procedures were regularly based on the calculation of probable or standard errors. The famous phrase ' more than three times the probable error ' (or ' twice the standard error '), which has figured so conspicuously in the literature of experimental psychology, was almost the hallmark of statistical analysis at that time. A classic instance, of special interest to the psychologist who of late has proved the chief user, is that of the tetrachoric correlation. This was but one of many coefficients devised by Karl Pearson for describing degree of association in a 2×2 table. For its probable error he proposed a complicated formula as long ago as 1913. Yet to the present day no exact sampling distribution has been furnished for a correlation so computed. Thus, not even if an underlying normal bivariate distribution is assumed do we possess any rigorous test of significance for this coefficient. Indeed, prior to 1915 the exact sampling distribution of the product-moment coefficient itself had not been determined, though both Gosset (' Student ') and Soper had given approximations.

Nevertheless, while the general problem of statistical inference remained confused, the line of progress had been indicated as early as 1908 when Gosset, in his classical paper ' On the Probable Error of the Mean ' (*Biometrika*, VI, pp. 1 *et seq.*), published the *t*-distribution. On the other hand, the theory of statistical estimation remained extremely backward. The importance of ' degrees of freedom ' was entirely unrecognized ; and thus the essential foundations of modern statistical analysis had still to be laid.

It was shortly after this that Professor Fisher made his first important contribution by solving the outstanding problem of the exact distribution of the correlation coefficient (*Biometrika*, X, 1915, pp. 507-21). This was followed a few years later by two articles in *Metron* ' on the probable error of a coefficient of correlation deduced from a small sample ' (which incidentally gave the exact sampling distribution of the intra-class coefficient) and on ' the distribution of the partial correlation coefficient.' The substance of these three articles forms the subject of the note which now makes up Paper I of the present volume : since " the established policy of *Metron* does not permit republication of articles from that journal," the papers themselves are not reproduced. The 1915 paper contained the first use of Professor Fisher's celebrated ' geometrical method ' (differing from earlier methods by the novel device of representing a sample of n variates as a single point)—a method which he brought to bear with such elegance and effect on various problems of sampling. In the second of

these articles we have the origins of the z -distribution and of the analysis of variance. In a biographical note (published in *Sankhyā*, 1938, and included in the introduction to the volume) Professor Mahalanobis discusses how Fisher's early training and practice in the use of geometrical methods may have been partly responsible for his subsequent applications of them; but he offers no clue to why Professor Fisher should have chosen to work with z , i.e., with half the natural logarithm of the variance ratio, rather than with the variance ratio itself, which was subsequently tabulated by Snedecor and labelled, by way of a pleasant compliment, with the initial F .

The papers which follow are presented for the most part in chronological order of publication. But here and there this is abandoned for a grouping according to subjects. The reasons for the mixed classification are not made clear. One would have thought that a subject classification should either have been used consistently throughout or else not introduced at all. As it is, some of the articles referred to in earlier sections do not appear till a later place in the volume, and no dates are given in the table of contents. Thus the subject of Paper 2 (1920), which contains the concept of a *sufficient* estimate, leads directly on to that of Papers 10 (1922) and 11 (1925), in which the problem of estimation is systematically discussed. The late insertion of Paper 10 is particularly confusing in view of the references to maximum likelihood that occur in earlier papers. In between we find Paper 3, which deals with a problem in the analysis of time-series and describes the use of orthogonal polynomials as "a means of isolating groups of components susceptible of separate explanation." This topic is then resumed in Papers 16 and 37 which deal with harmonic analysis. Incidentally we may observe psychologists, in spite of the frequent occurrence of time-series in researches on curves of work or learning, seem to have made little use of these methods.

Papers 4 to 9 have chi-squared as a common theme, all except the first being concerned with "goodness of fit." Before these articles were published it was the practice to take one less than the number of frequency groups as the quantity defining the appropriate chi-squared distribution for the test of significance, no matter what the form of the data. But here we are introduced to the now familiar concept of *degrees of freedom*. Paper 6 (1922) gives the distribution of regression coefficients, and is particularly rich in its implications. The t test is shown to provide the exact test for all regression coefficients 'linear or non-linear, simple or multiple,' and the z distribution, which, "before its general applicability was recognized, . . . kept turning up unexpectedly," is shown to be the appropriate distribution for testing the fit of regression lines, instead of the older chi-squared distribution. It is, however, unfortunate that the prefatory note to Paper 6 should read "Its [the z distribution's] relationships were first summarized (Paper 8) in 1924," when Paper 12, in the same year, seems evidently intended.

Paper 12 provides an admirable demonstration of the relations between the distributions of t , z , χ^2 , and a normal variate. The following paper (13) should also be of value to students, since it was written "to give mathematical teachers a concise account of the mathematical distributions used in the common tests of significance"; and the next paper (14) describes the general sampling distribution of the multiple correlation coefficient. Since very few statistical text-books written for psychologists offer a clear account of the relations between these different distributions, Professor Fisher's systematic discussions are especially to be recommended.

Papers 17 and 18 are concerned with problems which formed the starting point for Professor Fisher's book on *The Design of Experiments*; several later papers (39, 40, and 41) have to do with various mathematical ramifications of this subject. It is only during recent years that these principles have begun to influence the statistical and experimental procedures adopted (not always appropriately) by research workers in psychology.

Another series of papers, numbered 22, 25, 26, and 27, include the more important of his contributions to the problems of statistical inference. No one who employs tests of statistical significance can safely omit a proper study of the principles involved. The special features of Professor Fisher's treatment are, first, an unqualified rejection of Bayes' postulate, and, secondly, a strong emphasis on randomization rather than sampling as a condition for the valid use of statistical tests. However, even statisticians have found his discussion of these problems hard to follow; and the position is made no easier for the psychological research-worker by the presence of a rival approach (based on sampling and the theory of 'two types of error') which has been developed by Neyman and Pearson. The issues are primarily logical and may therefore in a large measure be appreciated without any advanced mathematical knowledge. Paper 27 (on 'Uncertain Inference') provides a comparatively easy approach: it examines the issues from an historical standpoint, contains no mathematics, and leads to a clear statement of Professor Fisher's own views. Of the remaining papers those of greatest interest to the psychologist are numbers 32, 33, and 34, which deal with the discriminant function—a device which has already been used in psychological investigations and has formed the subject of articles published in this *Journal*.

In view of its size and scope the volume can scarcely be considered expensive. It is primarily a work for the specialist. Yet it contains much that is of the highest value for all who aspire to an understanding of statistical methods rather than of mathematical statistics. *Arthur Summerfield*

Progress in Clinical Psychology. Vol. I, Section 2. Edited by D. Brower and L. E. Abt. New York : Grune & Stratton. 1952. Pp. xxiv + 266. \$5.00.

In their preface the editors announce that they have planned to issue every second year a volume summarizing the progress of clinical psychology. The first in the series, however, is intended to provide "as complete a coverage as possible of the past six years." Section 1, already published, presented "an historical and systematic résumé of developments" and a detailed account of "diagnostic and therapeutic instruments and procedures." Section 2 is divided into four parts, dealing respectively with developmental processes, applications to special areas, approaches to clinical psychology, and professional issues. Each chapter is written by a specialist in the branch, and includes a detailed appendix of references.

In studying the various summaries, the point that is most likely to impress readers of this *Journal* is the way in which the current developments in America are tending to stress the value of statistical research and particularly of quantitative checks on both theoretical and practical claims. As Professor Cattell observes, "American clinicians seem now to realize that their best hope of making substantial progress in clinical problems lies in the development of refined statistical procedures." The longest chapter in the volume is Professor Kogan's review of 'Recent Statistical Methods' as developed for clinical work. He deals with four main topics.

(i) First, he tells us, "one of the most active areas during recent years has been the attempt to devise systematic methods for handling profiles or score patterns in ways acceptable to the clinician." After reviewing the more important devices that have been proposed, he concludes that those likely to prove of the greatest service are the techniques of correlations between persons, of intra-class correlation, and of multiple discrimination. One interesting application of the principle of 'correlating persons,' the 'so-called Q-sort technique,' is being tried at the Counselling Center at the University of Chicago, suggested, it appears, by Dr. Stephenson : on the other hand, we are told, "Dr. Cattell has criticized both the nomenclature of the method and its dependence on non-objective data." Dr. Cattell himself contributes a separate chapter on P-technique. Intra-class correlation has the advantage of well-established tests of significance, but seems to have been used less freely in America than over here. In illustration of the possibilities of the discriminant functions Dr. Kogan refers to work reported in earlier articles in this *Journal*, particularly by Lubin, Rao, Slater, and others, in discriminating neurotic and normal groups.

(ii) A related field of work arises out of the need to develop special methods for validating tests of personality, since those hitherto used with tests of cognitive abilities have been dismissed by many clinical workers as inappropriate. Several investigators, notably Zubin, have put forward devices for validating projective techniques, while Cronbach has proposed a rather complex procedure (based on Vernon's blind matching procedure) for validating qualitative assessments of personality.

(iii) A third line of study that is attracting considerable attention is the attempt to develop criteria for assessing 'purity of measurement,' e.g., Guttman's method of scalogram analysis, Lazarsfeld's theory of latent structure, and the various factorial methods for studying the dominance of the general factor. Professor Kogan declares that he himself is "prejudiced towards prognosticating that the scaling approach (not necessarily linear scaling) will be found to meet the requirements more adequately than the approach by multiplex testing, but the scaled components must be synthesized back into reasonable models of human beings."

(iv) The 1952 symposium on 'Research Design in Clinical Psychology' (as well as other papers cited by Professor Kogan) has attracted increasing attention to the need for a more scientific planning of investigations in this field. Professor Kogan deplores the fact that so far he has been "unable to discover any comprehensive discussion of the logic of research designs comparable to that of Greenwood in the field of sociological research." The value of a control group has been increasingly recognized, especially in assessing the results of different modes of treatment. But he considers it disappointing that hitherto so little use has been made of Fisher's proposal to "substitute the multifactorial experiment for the single factor method to secure a broader basis for inductive generalizations." In Great Britain, 'largely owing to the direct influence of Professor Fisher himself, the 'latest refinements in the analysis of variance' have been more promptly and more widely used. "Recent developments," so Dr. Kogan believes, "portend a closer integration of analysis of variance and factor analysis." Here, however, he appears to have overlooked the work already done in this direction by Bartlett, Kendall, and those British factorists who have attempted to "integrate" the work of Fisher and of Pearson : as it happens, the very first issue of this *Journal* (1947, I, pp. 3-26) contained a systematic attempt to show the mathematical relations between analysis of variance on the one hand and analysis of covariance on the other. Dr. Kogan rightly points out that clinical investigators are far too ready to declare that their most pressing problems are unamenable to the stock statistical techniques, without considering whether some of the newer devices may not furnish the very method which their problem demands. Among more recent developments the merits of

so-called "non-parametric or distribution-free methods of statistical inference," such as the 'order statistics' of Wilks and the ranking methods of Kendall, rightly receive considerable stress.

In his chapter on 'P-technique Factorization,' Professor Cattell begins by urging a greater use of statistical methods generally. For clinical research he, too, considers the most useful to be factor analysis and the analysis of variance, since they do not require that artificial control of the variables, or that artificial elimination of all variables except one, which has proved so difficult to achieve in dealing with human beings. Factor analysis he regards as "the ideal device for dealing with the total personality"; and in particular, he maintains, "the new factorial procedure known as P-technique¹ gives precisely an analysis and description of unique personal traits."

In the remaining chapters the surveys of the literature on special branches of clinical psychology may enable the reader to judge the uses made of statistical methods in this or that field by the summaries they give of the latest results. Their general effect, it may be said, is to demonstrate the need for greater caution in regard to the somewhat optimistic claims so often put forward in the field of clinical psychology during the last few decades.

The publication of these biennial surveys will be welcomed in this country as warmly as in the United States. It is therefore unfortunate that no very systematic attempt seems to have been made to cover recent British contributions, especially as several of the methods and results quoted as new by the American authors have long been accepted over here. There are generous references to work published in this *Journal*, but practically none to relevant articles in the other psychological journals published in England. Nevertheless, in their selection of material and in their appraisal of results, the general editors and the authors of the several chapters display a discernment that is at once sound, critical, and impartial. Certainly, if we may judge by the two sections now issued, the series will be of the greatest value to all students of clinical psychology, and should be a challenge to the statistical psychologist to produce and apply more appropriate research-techniques in this important field of work.

Cyril Burt

The Young Delinquent in his Social Setting. By T. Ferguson. Oxford University Press, 1952. Pp. xii + 158. 10s. 6d.

In an earlier volume (*The Young Wage-Earner*, 1951) Professor Ferguson has described an enquiry into the after-histories of 1,349 Glasgow boys who left the ordinary schools at 14. The present study forms a supplement to that broader survey. Its object is to determine the frequency of juvenile delinquency and to ascertain its apparent causes. With a view to obtaining further light on physical and mental factors, minor studies were also made of smaller groups from schools for the physically or mentally handicapped.

He finds that over 12 per cent. of the boys from the ordinary schools had been convicted of crime before the age of 18. Among these "a low level of scholastic ability was probably the most powerful single factor associated with a high incidence of crime." Unfortunately it was not possible to test their intelligence; and the reader may wonder whether the real factor was, not so much the low level of scholastic attainments, but rather the low level of innate capacity, which, as earlier enquiries have shown, is a common characteristic of juvenile delinquents and the main cause of educational backwardness. No attempt was made to assess emotional or temperamental qualities or to determine whether these may not be even more important than intellectual.

In common with most psychological investigators, Professor Ferguson comes to the conclusion that physical factors by themselves have little influence, but that environmental conditions play a considerable part. Of these the most important, he believes, are frequent change of employment and residence in an overcrowded district. The influence of 'broken homes,' he finds, is not so great as other writers have supposed: "in good housing districts the proportion convicted was below the average, whether the family was 'broken' or not; in bad housing districts the proportion was above average, and slightly worse in families which were *not* 'broken.'" Among those who were physically handicapped "the absence of a parent" proved less important than "the presence of a bad parent."

If a youth lives in a home with three or more persons to a room the risk of his becoming delinquent is about five times as great as if he lived in a room with less than three. Professor Ferguson infers that "the abatement of crowding of a severity of three persons or more to a room would apparently produce a reduction in the amount of juvenile delinquency equal to about one-fifth—provided that in

¹The term is used to denote 'the factorization of an individual case' (i.e., what I myself had referred to as 'intra-individual variability': cf. Burt, C., 'Correlation between Persons,' *Brit. J. Psych.*, XXVIII, 1937, p. 60). Professor Cattell explains that there is a tendency to extend Stephenson's term 'Q-technique' to cover correlations between *different* persons, and therefore "it would be convenient to adopt Burt's term P-technique to refer to this analysis, within a *single* person," i.e., what he himself had elsewhere called 'intra-individual covariation' (cf. Cattell, R., *Description and Measurement of Personality*, p. 99).

their new setting the families conform to standards of conduct prevailing among families not at present severely crowded." But will they conform? That surely is the crux. By the title he has chosen Professor Ferguson tacitly invites a comparison between his methods and results and those set forth in the volume on *The Young Delinquent* by Professor Burt. Now Burt has shown that, to a very large extent, delinquency is a by-product of a general inferiority, characterizing families at the lower end of the socio-psychological scale and apparently resting on an innate and inheritable basis. Intellectual, educational, temperamental, moral, and civic inferiority are all correlated together; and such families tend to create or drift into the unhealthy overcrowded districts where delinquents are commonly found. Improvement in living conditions might do something towards solving the problem, particularly in the case of those individuals in whom extrinsic conditions are more important than intrinsic; but it could hardly cut down delinquency to the extent he implies.

Professor Ferguson's conclusions might be summed up by saying that, so far as the conditions covered by his survey are concerned, the factors conducing to youthful delinquency appear after all to be much the same in Glasgow as in London and other large cities where surveys have already been made. It is, however, not altogether easy to decide how far his data really support the inferences that he has drawn. To begin with, he ignores the somewhat special conditions which, according to other social investigators, appear to prevail in Glasgow. Burt, in his early studies at Liverpool, found that the amount of delinquency among children of Irish immigrants or of Catholic parents was disproportionately high; and this was confirmed in a later study by Bagehot. It is conceivable that a similar conclusion might hold good if the Glasgow figures could be analysed according to districts. Professor Ferguson prints figures for 'Church attendance,' but says nothing about religious denomination. And in general "the sociological information which he gives is" (as another critic has remarked) "woefully inadequate."

Medical writers who deal with psychological problems are fond of criticizing "the psychologist's passion for statistical calculations." Accordingly, one of Professor Ferguson's chief aims was "to bring out the relations (between crime and causal factors) *without* invoking the aid of elaborate statistical techniques." Thus, when a difference is observable in the figures obtained from two contrasted groups, he appears to accept it straight away if it agrees with his preconceptions (even though it is not statistically significant) and to ignore it if it does not. No endeavour is made to use the figures from the larger control group to eliminate the influence of the various conditions as found among the delinquent groups (e.g., by calculating coefficients of contingency or correlation). To demonstrate the effects of subsidiary factors all that is done is to omit those cases in which the major factor is found, and then recalculate the percentages—a somewhat fallacious procedure.

In criticizing the methods adopted by Professor Ferguson in his earlier survey, Mr. Alec Rodger declares that the writer "has slipped on practically every banana-skin there was for him to slip on";¹ and he agrees with the critic in *The Times* who complained that, in their discussions of delinquency, the investigators had constantly confused mere association with actual causation. The present volume endeavours, not (it must be confessed) with very great success, to circumvent these objections. As Mr. Rodger points out, many of the difficulties have arisen because, from the very outset, the essential problems were carelessly formulated in terms very loosely defined, and the whole enquiry as a result was somewhat amateurishly planned. Nevertheless, the tables contain much useful material. There is an obvious moral to be drawn: if those who are experts in medicine wish to launch out on enquiries in the psychological and sociological fields, they should be urged to seek the co-operation, or at least the advice, of those who are familiar with the innumerable pitfalls and trained in the appropriate techniques.

H. L. Jackson

Contributions of Psychology to Social Problems. By Cyril Burt. London: Oxford University Press. 1953. Pp. 76. 5s.

At the beginning of this century, L. T. Hobhouse, the first Professor of Sociology at the London School of Economics and Political Science, was looked upon as the leader in what was then the newly christened field of social psychology. The Webbs, to whom the establishment of the School was chiefly due, had been eager to replace the older, speculative brand of political theory and practice by "a scientific discipline founded on objective methods of research," particularly that of "quantitative study based on first-hand data," which had been introduced by Charles Booth and Francis Galton. "Every conclusion," said Mrs. Webb, "which is derived from observation or experiment, has to be qualified, as well as verified, by relevant statistics."² Accordingly, social psychologists like Hobhouse, Graham Wallas, and the young McDougall—all working together at that time in London—set themselves to formulate, in the shape of tentative hypotheses, the basic problems which appeared to confront the scientific investigator in the new field of work.

¹ *Occupational Psychology*, XXVI, 1952, pp. 251-2.

² Beatrice Webb, *My Apprenticeship*, p. 340.

In his present survey, which forms the substance of his Hobhouse Memorial Lecture, Sir Cyril Burt seeks to summarize the chief advances made during the last fifty years in the many-sided attempts to solve such questions. One of his incidental aims, he tells us, has been to draw attention to British contributions, "partly because recent writers (who appear more familiar with American work) have so often ignored them, partly because it seems rash to assume that conclusions about social matters reached by investigators in other countries will also hold good over here," but above all because he believes that "during and since the recent war British psychologists, while devoting themselves more and more to practical applications, have tended increasingly to neglect research on fundamental issues." He points out that, although research in social psychology has been carried out far more extensively in America than over here particularly during the last two decades, nevertheless, from the time of Booth's surveys onwards, such work has exercised a far more direct effect on legislative and administrative changes in this country than in the United States.

He begins with a brief account of the special scientific techniques—observational, experimental and statistical—devised for such investigations. Most of them, he claims, owe their origin to British pioneers—Galton, Booth, Pearson, the Webbs, and Hobhouse himself. To a large extent, as he points out, they were first tried and refined in the course of investigations on social and educational enquiries conducted with the aid of the schools, where material was more readily accessible.

Turning to the results hitherto attained by the use of such procedures, he considers, in the first part of his review, what are the special mental characteristics of the human race that specially affect social organization and social relations. He concludes that they are chiefly 'similarities in the so-called social instincts' and 'differences in individual ability and temperament.' In the second part he discusses, in somewhat greater detail, the converse problem, namely, the influence of social organization and relations on the mental characteristics of individuals, as they develop in a civilized community. Here he recapitulates the main results achieved by statistical and experimental methods on the problems of sex, social class, and social attitudes and opinions.

The third part summarizes a number of typical enquiries, again based mainly on the combination of case-studies and statistics, in regard to recent trends and changes in religion, morality, leisure occupations, and the like. The whole survey forms a well-packed yet highly readable account of the conclusions so far established, and closes with a strong plea for systematic research in the social branches of psychology comparable to that which has achieved such striking success in the educational branches.

J. S. Mason

The Causes and Treatment of Backwardness. By Cyril Burt. London: National Children's Home, 1952. Pp. 144. 4s. 6d.

This is the most recent of a series of books, published at a surprisingly low price by the National Children's Home, and dealing in non-technical language with problems in child psychology and education. The purpose of the present volume is to explain to the teacher, the parent, and the student of other branches of psychology the need for a scientific investigation of educational sub-normality, and to present a systematic summary of the main conclusions hitherto reached. The author lays special stress on the aid given by quantitative procedures, like intelligence-tests and standardized tests for special aptitudes and scholastic attainment, and above all on the importance of a statistical evaluation of the available evidence in order to check impressionistic views about causation and treatment such as general experience or theoretical knowledge may suggest. At the same time he emphasizes still more forcibly his view that tests and statistics by themselves are quite inadequate.

For the practical investigation of individual pupils he proposes a scheme of case-study, and contrasts it with the common type of case-report submitted by teachers and even clinical psychologists, where little or no order is observed either in the investigation or in the written account, and where facts, inferences, personal impressions, and favourite dogmas are all confused together. The character sketches that he himself includes provide vivid descriptions of typical individuals, and demonstrate that it is quite possible to combine human and sympathetic insight with a demand for precise measurement and patient statistical verification. The dry bones live. Incidentally, useful suggestions are made as to the ways in which teachers, particularly those with a professional knowledge of mathematical procedures, might contribute to the advancement of the subject by carrying out researches in their own schools. The practical recommendations on the organization of special schools, special classes, and special remedial schemes suited to each individual case make this work an admirable manual for the beginner in clinical psychology. For the statistical psychologist its chief interest is perhaps to demonstrate how abstract quantitative procedures need not be incompatible with a concrete interest in human personalities.

E. A. Gardner

Facts from Figures. By M. J. Moroney. London : Penguin Books, Ltd. 1951. Pp. viii + 472. 5s.

Mr. Moroney describes his book as "an attempt to take the reader on a conducted tour of the statistician's workshop": the visitor is allowed to watch experts at their task, and then encouraged to try his own hand. Those who are invited are "not only students, but all whose work calls for a general knowledge of the subject, particularly in the world of industry and research." And they are assured that "the rapid expansion and enormous success of statistical techniques in recent years is ample proof of the need for such methods."

The publishers are certainly to be congratulated on producing a book of nearly 500 pages with diagrams, numerous tables, and formulæ that must have cost the compositor some skill—all for the price of two half-crowns. The introductory explanations are racy, lucid, and entertaining. Indeed, every chapter contains ingenious analogies or illustrations which will often prove instructive, not only to the novice, but even to those already acquainted with the stock procedures. Special attention is paid to those branches that are applicable to industrial production and quality control. The more advanced sections are, on the whole, fairly up to date; and the concluding chapters deal with some of the most recent devices in the analysis of covariance, the study of time series, and the use of ranking.

On the other hand, many pages betray signs of haste or carelessness. The theoretical explanations are often inadequate, and some would certainly be frowned upon by statistical critics. There are awkward gaps in the arguments, references to a diagram which does not exist, and numerous misprints, in addition to those contained in the publishers' *errata* (which relate only to four out of the twenty chapters). Some of the distributions are badly grouped, some of the percentages wrongly calculated; and several of the answers to questions given at the end are incorrect. No doubt most of the slips will be corrected in a later edition. Meanwhile, as a semi-popular introduction the book should be of great help to those students who are unable to afford the most recent text-books or who find the more academic expositions of mathematical procedures too dull, too abstract, or positively bewildering.

Cyril Burt

